BEYOND FRAME SHUFFLING: INTERPRETING TEMPORAL-ORDER SENSITIVITY IN VIDEO PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A video predictor's response to reordered frames is difficult to interpret: reordering can change local transitions and perturbation magnitude, while aggregate error scores can conceal changes in the predictions themselves. We study this ambiguity through controlled comparisons with explicitly preserved input properties. Our construction enumerates six histories with identical frame content, boundary frames, and final five observations, and groups them by their directed adjacent-frame-pair multisets. Within-group pairs preserve local transitions; between-group pairs provide references with the same frame content. We evaluate all pairs using prediction disagreement, pixelwise error change, and aggregate error change. Experiments comprise 896 forecasts on 32 Moving MNIST clips under unmasked and masked observations, using released SimVP and OpenSTL ConvLSTM predictors. For unmasked first-frame predictions, SimVP's transition-matched pairs exhibit 14.7% greater mean prediction disagreement than unmatched pairs. Post-hoc comparison against references with exactly equal input L_1 distance reduces this gap to 6.3%, with a similar result under stricter per-pixel distance matching. The ConvLSTM predictor remains close to parity, and some masked comparisons reverse the ordering. Output-range decomposition and stepwise analysis further distinguish prediction changes from those retained by scalar errors. Together, the controls and decompositions turn order sensitivity into a precisely specified comparison: the input constraints define what is held fixed, and the readout defines which prediction changes remain visible.

1 INTRODUCTION

Changing a video's frame order can change a predictor's output. Interpreting that response requires a comparison: how much change should remain when frame content, local transitions and perturbation magnitude are controlled? Even after these input properties are fixed, the evaluator must choose which eligible reference pairs receive weight. We study whether those remaining choices determine the reported response direction and the ordering of predictors.

Our central question is: *which conclusions are identified by the declared controls, and which additionally depend on the reference rule?* We distinguish a score at one rule from a direction that holds for every rule satisfying the same constraints. Endpoint balance controls the contribution of individual histories; pairwise prediction responses can still depend on how those histories are compared. Optimizing the contrast over the declared reference domain makes that distinction testable on unchanged forecasts.

The initial Moving MNIST study motivates this question. Across 896 forecasts on 32 clips, SimVP's unmasked first-frame contrast is 14.7% with all references and 6.3% after post-hoc distance matching. A prospective 32-clip follow-up does not confirm the positive first-frame direction: ratios are 0.9996 for the original construction and 0.8978 for a second construction. Rather than treating the initial sign as a predictor property, we examine which comparison choices sustain it.

The evidence follows one comparison from reference choice to replication; a separate analysis connects output response to task consequences. A direct common-weight comparison reverses SimVP and ConvLSTM on first-frame KTH probes, while comparisons involving PredRNN remain ordered

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

throughout the half-floor domain. Freezing the two reversing weight vectors before acquiring 600 forecasts on thirty additional segments from the same five subjects reproduces the first-frame reversal in all five subjects, including after model-specific normalization. A separate same-perturbation analysis at the mean-ten horizon asks when output response corresponds to a loss of forecast quality: in the retrospective screen, the PredRNN/ConvLSTM ordering agrees on both measures in four of five subjects and after every subject deletion. Fixing this selected pair before acquiring 720 forecasts on later, nonoverlapping windows reproduces the relation in all five subjects and every deletion. This task check uses native-to-perturbed response, distinct from the matched/reference contrast.

The extension to histories that retain every frame once admits 15/30 clips and yields primary range crossings on 11/15. It transfers an input-defined comparison algorithm using a transition-magnitude descriptor, rather than the original directed-transition multiset. In both studies, agreement among three graph-only central rules coexists with wider-domain ambiguity. We therefore report the chosen rule, the admissible range and the distance to a ranking boundary together: central-rule agreement, local stability and domain-wide invariance describe different properties.

The contribution is a controlled account of reference-dependent interpretation: which comparisons remain open after input matching, which directions survive fixed-reference transfer, and which response orderings have an observed task consequence. The full floor path and local-radius checks specify the assumptions behind each conclusion. Pooled accuracy rankings remain stable and model-selection gains fail deletion checks, separating those applications from the measured response effects.

2 Defining the Comparison

Let a history contain frames indexed by $N = [0, 1, \ldots, 9]$. For each permutation (a, b, c) of $(0, 1, 2)$, family A contains

$$P_{abc} = [0, a, b, c, 0, 5, 6, 7, 8, 9]. \tag{1}$$

These six histories share content, boundary frames and their final five observations. They form three directed adjacent-pair multiset classes: $\{P_{012}, P_{120}\}$, $\{P_{021}, P_{210}\}$ and $\{P_{102}, P_{201}\}$. A within-class pair preserves every position-independent additive directed-pair summary,

$$\sum_{t=0}^{8} \phi(x_t^A, x_{t+1}^A) = \sum_{t=0}^{8} \phi(x_t^B, x_{t+1}^B). \tag{2}$$

Temporal assignment remains free. The identity follows from multiset equality and specifies the controlled class of input statistics. We call within-class pairs matched and cross-class pairs references. Actual pixel equality, checked after any masking, determines the realized classes; a nominal distinction can disappear when frames coincide.

Family B uses $E_{abc} = [0, a, b, c, 0, 0, 6, 7, 8, 9]$ and $L_{abc} = [0, 0, a, b, c, 0, 6, 7, 8, 9]$. Two aliases leave ten unique histories, with class sizes 3, 3 and 4. Content and the last four frames agree within B. Both content multiplicity and suffix length differ between A and B, so B tests a second construction rather than isolating a recency intervention. Error readouts use the original clip's future as the target. These constructed histories use repeated and omitted frames to control the specified statistics.

Reference support. For a matched pair m, retain references r with the same input L1 distance. A stricter rule equates the full spatial map of time-summed absolute input differences. Denote the eligible joint pairs by $E = \{(m, r)\}$. These two rules coincide on the prospective Moving MNIST slice, so they do not provide independent replications there. The initial slice has exceptions; all are retained in the supplement. An empty reference set makes the requested comparison undefined, rather than zero.

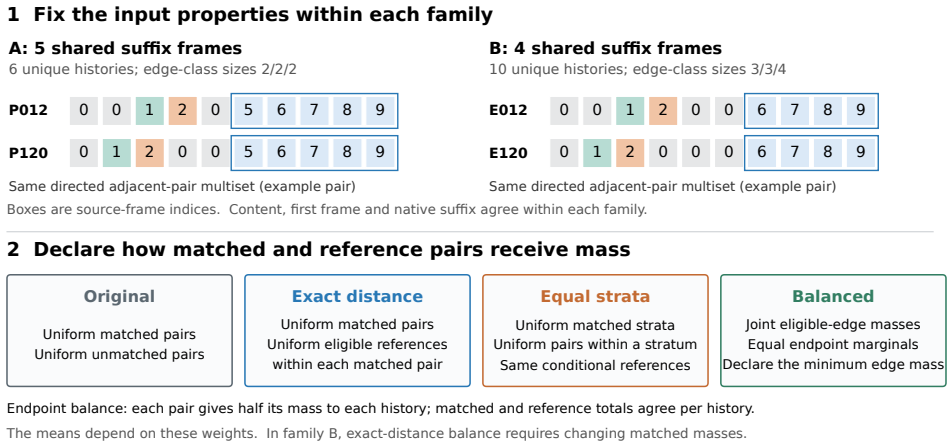

Figure 1: A controlled probe needs both histories and a comparison rule. The top shows representative matched histories from the two families. The bottom separates reference support from weighting choices. Exact distance constrains eligible pairs; endpoint balance constrains history marginals, while the weights still determine which pair interactions contribute. Content is preserved within each family, not between families.

Readout. For n pixels, let $e_{A,i} = |z_{A,i} - y_i|$ and $e_{B,i} = |z_{B,i} - y_i|$. Prediction disagreement, pixelwise error change and aggregate error change are

$$Q = \frac{1}{n}\sum_i |z_{B,i} - z_{A,i}|, \qquad L = \frac{1}{n}\sum_i |e_{B,i} - e_{A,i}|, \tag{3}$$

$$D = \left|\frac{1}{n}\sum_i (e_{B,i} - e_{A,i})\right|. \tag{4}$$

Thus L takes absolute values *before* averaging, whereas D takes one absolute value *after*. The triangle inequality gives $D \le L$; the pixelwise reverse triangle inequality $|e_{B,i} - e_{A,i}| \le |z_{B,i} - z_{A,i}|$ gives $L \le Q$.

For two pixels, take $y = (.5, .5)$, $z_A = (.1, .6)$ and $z_B = (.7, .8)$. Errors change from $(.4, .1)$ to $(.2, .3)$: opposite changes cancel in $D = 0$, while $L = .2$ and $Q = .4$. An unchanged scalar error can therefore hide both error redistribution and changed predictions.

Response magnitude and forecast accuracy answer different questions. Our primary comparison uses raw first-future Q; the ten-frame mean is a separate readout. Equal-weight clip means define each aggregate, with no pseudoreplication from overlapping pairs. A predictor using only the shared suffix gives zero response within either family, so zero response alone cannot distinguish useful invariance from prefix insensitivity.

3 Feasible References and Direction Ranges

When can reference weights preserve a comparison? Let a_s be the incidence vector of a pair's two histories. Joint edge masses w_{mr} induce matched masses $u_m = \sum_r w_{mr}$ and reference masses $v_r = \sum_m w_{mr}$. Endpoint balance requires

$$\sum_{(m,r)\in E} w_{mr}(a_m - a_r) = 0. \tag{5}$$

This is exactly the condition that cancels every additive single-history contribution: for any history scores g, the induced contrast is $g^\top \sum w_{mr}(a_m - a_r)$, which vanishes for all g if and only if balance holds. Thus balance removes unequal endpoint composition without specifying the pairwise response Q.

Preserving given matched masses adds row-sum constraints on these joint weights. A convex-separation test decides whether the reference side can match the required endpoint marginals; the full conditions and proofs are in Appendix S24. In B, equal matched-pair masses are infeasible under endpoint balance, whereas flexible joint weights can balance the graph. This changes the matched comparison as well as the reference and must be reported as a change of estimand.

A range, rather than one selected balanced weighting. For a declared floor τ, define

$$\mathcal{P}_\tau = \{w : \mathbf{1}^\top w = 1, \sum_{(m,r)\in E} w_{mr}(a_m - a_r) = 0,\ w_{mr} \geq \tau\}. \tag{6}$$

We parameterize $\tau = \alpha t_*$, $0 \leq \alpha \leq 1$, where t_* is the largest feasible minimum edge mass ($1/12$ for A, $1/88$ for B). Zero floor allows every balanced nonnegative weighting; increasing α requires more mass on every eligible edge. The earlier half-floor is one slice of this nested sensitivity path, not an empirically preferred distribution. For clip i, write $c_{i,mr} = Q_{i,m} - Q_{i,r}$. Shared weights give

$$[\ell_s, h_s] = \left[\min_{w\in\mathcal{P}_\tau} \bar{c}^\top w,\ \max_{w\in\mathcal{P}_\tau} \bar{c}^\top w\right], \qquad \bar{c} = \frac{1}{n}\sum_i c_i. \tag{7}$$

Clip-dependent weights instead average the per-clip extrema, yielding $[\ell_p, h_p]$. The shared domain embeds diagonally into the product domain, hence $\ell_p \leq \ell_s \leq h_s \leq h_p$. For a ten-frame readout, each cost is time-averaged before optimization; the same weights apply across prediction steps. Averaging independently optimized step endpoints would answer a different question.

For comparison with the initial study, we also report the ratio

$$\rho_Q(w) = \frac{\sum_{(m,r)\in E} w_{mr}\bar{Q}_m}{\sum_{(m,r)\in E} w_{mr}\bar{Q}_r},$$

with strictly positive denominators verified over each domain. Table 50 gives minima and maxima of this ratio under shared weights; Table 1 instead gives signed-difference ranges under both shared and product weights. A ratio above one and a positive difference have the same direction for a given weighting, but their extrema need not use the same weights.

What the evaluator can report. A positive lower bound supports a positive contrast for every weight in the declared domain; a negative upper bound supports a negative one. Bounds on both sides of zero exhibit admissible weights with opposite conclusions on the same outputs. These are finite-domain statements, not confidence intervals. Outcome-optimized endpoints diagnose sensitivity and are not newly validated weighting policies. Rational primal/dual certificates verify extrema relative to the stored numerical costs. The saved-output entry point reproduces the shared/product table without forecasting again; its inputs, witnesses and certificates accompany the supplement.

For the prospective comparison, a fixed subset of endpoint-balance rows also yields a sufficient direction certificate through the local LP reduced costs (Appendix S10.1). We evaluate its retained directions rather than treating its mathematical soundness as empirical success. It still requires local optimization and is not proposed as a faster replacement for the full interval.

4 Empirical Evaluation

Scope of the evidence. The initial 32-clip Moving MNIST study found a positive ordering; the prospective follow-up produced different directions (ratios 0.9996 and 0.8978). The studies below test whether controls identify a direction, whether a common rule identifies a model ordering, and whether the same reversing weights transfer.

Study design. We use released OpenSTL SimVP+gSTA, ConvLSTM and PredRNN artifacts (Tan et al., 2023; Li et al., 2023) on Moving MNIST (Srivastava et al., 2015) and KTH (Schüldt et al., 2004), with ten history frames and ten retained predictions. Our experiments perform no training or checkpoint selection; the upstream test-as-validation protocol limits selection independence. These are named artifacts, not samples of architecture families. The earlier four-subject KTH cohort admits 18/24 clips after pixel-equivalence checks. The supplement retains implementations, complete forecast grids and exclusions.

4.1 INPUT MATCHING AND BALANCE LEAVE THE DIRECTION OPEN

Table 39 compares the initial observation and prospective test. Matching also changes reference weights, so its reduction does not isolate distance causally. Prospective A fails the positive criterion (0.9996), without establishing equivalence. In B, exact-distance references change 1.0659 to 0.8978 on identical outputs; every delete-one-clip ratio stays below one. Secondary A at mean ten (1.0209) does not rescue its failed primary criterion.

The initial ConvLSTM comparison is near parity: its unmasked first-frame Q ratio is 0.9978. The initial mean-ten ratio changes from 0.9969 unmasked to 1.0008 masked, a small ordering reversal rather than evidence of a substantial effect or equivalence. These results supply the masked and near-parity comparisons noted in the abstract; complete initial-arm readouts remain in Appendix S19.

Balanced reference domains. On the Moving MNIST follow-up, A permits balance with equal matched masses; B requires changing those masses. Both primary SimVP shared-weight ranges cross one with positive mass on every eligible edge (Table 50). At the maximum possible minimum edge mass, A is fixed below one and B is wholly below one: the additional weight objective determines that direction.

The retained grid has 1,728 envelope rows: 576 correlated settings of artifact, arm, family, horizon and readout across three floors. Both one-sided and crossing intervals occur. Appendix S24 distinguishes fixed matched masses, flexible balance and positive support.

4.2 A COMMON WEIGHTING RULE SEPARATES TWO KINDS OF AMBIGUITY

For KTH, all 18 supported clips share identical named edges and balance equations within each family. This permits a direct shared/product-domain comparison with no change of cohort or edge support. Table 1 reports all eight model/family/horizon cells. The shared-weight analysis is post hoc; it adds no forecasts.

Table 1: All KTH model/family/horizon cells on the same 18/24 conditional clips. Signed $Q_M - Q_R$ bounds are scaled by 10^5. Product weights may vary by clip; shared weights may not. The half-floor is identical. Endpoints are exact relative to archived binary64 costs, displayed rounded.

Artifact	Family	Readout	Product weights	Shared weights
SimVP	A	first	[6.729, 8.221]	[6.764, 8.186]
SimVP	A	mean 10	[7.364, 9.287]	[7.769, 8.882]
SimVP	B	first	[-4.107, 6.579]	[-3.996, 6.512]
SimVP	B	mean 10	[-4.656, 9.837]	[-4.078, 9.158]
ConvLSTM	A	first	[0.023, 1.671]	[0.023, 1.671]
ConvLSTM	A	mean 10	[-2.284, -0.239]	[-2.267, -0.256]
ConvLSTM	B	first	[-4.022, 2.155]	[-3.848, 2.083]
ConvLSTM	B	mean 10	[-10.579, 1.109]	[-9.017, -0.313]

First-frame B spans both signs on both artifacts even with one common weight vector. Its ambiguity therefore does not require adapting weights to individual clips. A is positive throughout both first-frame domains; shared positivity follows mathematically from product positivity and is not another replication. For ConvLSTM A at the first frame, shared-feasible weights attain both product extrema, so the intervals coincide. Its lower endpoint is only 2.329×10^{-7} in raw units, establishing finite-domain sign rather than a practical robustness margin.

ConvLSTM B at mean ten supplies the complementary result. Its product interval spans zero, but its shared interval is entirely negative. Obtaining a positive contrast within the declared product domain requires relaxing the common-weight constraint. The evaluation conclusion changes from direction-ambiguous to uniformly negative when the probe must use one common rule. SimVP B remains two-sided at mean ten. The sharing convention now matches the Moving MNIST analysis, but different graphs, datasets and checkpoints still prevent a controlled attribution of their differences.

Earlier horizon dependence and aggregation loss appear in Appendix S11.1.

The five-subject KTH cohort contains 21/30 eligible clips. Its third predictor, PredRNN, is negative throughout all twenty subject and four pooled half-floor cells, while some SimVP and ConvLSTM cells span zero (Appendix S13.1). This motivates distinguishing each predictor's own direction from their relative ordering.

4.3 Reference choices reverse a cross-predictor ordering

We compare predictors by optimizing $(\bar{c}_f - \bar{c}_g)^\top w$ under one common weight vector. In this post-hoc saved-output analysis, SimVP minus ConvLSTM on first-frame B spans $[-2.207, 3.940] \times 10^{-5}$ at the half-floor over 21 eligible clips. Both endpoint witnesses preserve the graph, balance and positive edge floor. The lower-endpoint weights rank ConvLSTM higher; the upper-endpoint weights rank SimVP higher (Figure 2).

Over the half-floor domain, the pooled gap spans $[-7.47\%, 13.33\%]$ relative to the fixed input-defined-witness scale. All five per-subject intervals of this cell cross zero, with more than 5% of each subject's scale on both sides. The same pooled witness pair reverses all five raw and model-normalized orderings. At these two witnesses, the models' relative contrasts $C_f/[(M_f + R_f)/2]$ also reverse: pooled SimVP/ConvLSTM scores are $-5.10\%/+0.56\%$ and $+7.20\%/-6.41\%$, respectively.

Across three model pairs, 9/60 correlated subject comparisons and 1/12 pooled comparisons cross zero at the half-floor; the others stay fixed. All 40 subject and eight pooled comparisons involving PredRNN stay positive despite the other models' first-frame B ambiguities: per-model direction ambiguity need not change model order. At the maximum floor all raw orderings are fixed. Appendix S13.2 reports pooled ranges and magnitude diagnostics; all 216 ranges accompany the replay. Accuracy rankings require separate outcomes.

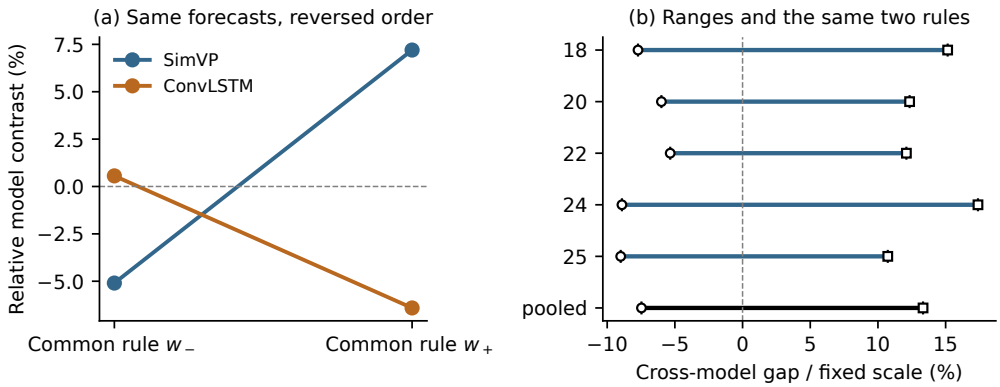

Figure 2: A common reference rule changes the ordering of SimVP and ConvLSTM on first-frame B. (a) Both predictors use the same two admissible weight vectors; relative contrasts normalize each model by its own average matched/reference response at that weight. (b) Exact raw-gap ranges divided by the fixed witness scale, separately for each subject and the pooled cohort. Circles/squares apply the same pooled w_-/w_+ to every subject. These are reference-choice ranges, with five subject clusters, rather than sampling intervals.

4.4 A frozen reference pair reproduces the ranking reversal

Do the same reversing weights transfer? Unchanged first-frame B witnesses w_-, w_+ (Section 4.3) reverse raw and normalized rankings in all four earlier subjects and their pool. At mean ten, only one subject reverses and the pool does not. Prospectively, the second annotated sequences of the same thirty videos all pass the pixel-graph test. Their 300 inputs are new; two frozen predictors yield 600 forecasts with no weight optimization.

New-sequence first-frame rankings reverse in the pool and all five subjects, raw and normalized (Table 2). These are new segments of the same subjects and videos. At mean ten, neither pooled comparison reverses; only 2/5 subjects reverse raw and none normalized (Appendix S13.3).

Table 2: Unchanged first-frame B witness weights on three cohorts; parentheses give admitted clip counts. Raw gaps are SimVP minus ConvLSTM in units of 10^{-5}; normalized gaps are percentage points. New-sequence weights and eligibility are fixed before forecasting.

Cohort	Raw w_-	Raw w_+	Norm. w_-	Norm. w_+
Weight-selection (21)	-2.207	3.940	-5.66	13.61
Earlier cohort (18)	-3.228	6.801	-4.56	16.06
New sequences (30)	-1.681	3.407	-6.68	16.19

Input-defined weighting choices. We fix three graph-only objectives: maximum entropy, minimum squared edge mass on $\mathcal{P}_{t_*/2}$, and maximum minimum mass with a minimum-squared-mass tie-break. Their weights use no forecast values; they are mathematical baselines for reference choice. On the saved three-predictor grid, the three rules disagree on the raw ordering in only 1/60 subject comparisons and 0/12 pooled comparisons. After model-specific normalization, all three agree in every comparison. The lone raw disagreement is SimVP–ConvLSTM, B/first, subject 25; its gaps span $[-1.260, 2.707] \times 10^{-6}$. Thus the full-domain reversal does not imply that these central weighting choices commonly disagree.

4.5 Reference ambiguity with every frame retained once

We test a new construction on thirty d2 videos from the same five subjects. Its 24 anchored permutations use every original frame exactly once. Matched pairs share a histogram of adjacent-frame mean pixel-L_1 magnitudes at width $1/255$, fixed on d1 inputs; joint edges preserve exact input-distance maps and endpoint balance. This descriptor replaces directed-transition matching. Clip-specific input graphs admit 15/30 candidates; the transferred object is the input-defined rule, not identical numerical weights. Three unchanged predictors yield 1,125 remote forecasts (Appendix S13.4).

For first-frame SimVP–ConvLSTM comparisons, the full raw-gap range crosses zero on 11/15 clips, covering all five subjects. The three central rules disagree on only one raw ranking and agree on every normalized ranking. Thus reference-domain ambiguity persists without repeated or omitted frames, while agreement among these rules persists as well.

4.6 When a reference-sensitive report is informative

Agreement and local stability. The full range answers whether all admissible choices share a direction; it does not measure how close a chosen rule is to reversal. On the eleven content-preserving primary comparisons admitting reversal, median total-variation distances to the zero boundary are 14.40%, 14.39% and 10.25% for entropy, squared-mass and maximum-floor centers. A 5% budget permits reversal on 3/15, 4/15 and 2/15 clips. Thus central-rule agreement and local stability can accompany a two-sided full range (Appendix S13.5).

Reporting under a specified domain. On five new KTH subjects, a point-margin threshold calibrated on an earlier cohort ties full-range reporting at the first frame: each retains 10/20 directions. At mean ten, the threshold reports eleven directions, including two whose full ranges are two-sided; the interval supports ten. The comparison concerns extending a point sign to the specified domain. Its reference standard is domain invariance, rather than an external label of temporal ability. Post-hoc calibration variants and full subject results remain in Appendix S10.

Dependence on the floor. In the five-subject, two-artifact pixel-disagreement grid, 17/40 cells are strictly one-sided at zero floor and therefore throughout the nested path. Twenty additional cells require a positive floor, including three already one-sided at half-floor; three remain non-one-sided at maximum floor. These correlated finite comparisons separate directions requiring no minimum mass from those sustained by that extra assumption (Appendix S11). Strict means $\ell > 0$ or $h < 0$; touching zero does not qualify. Changing the readout to $1 - \mathrm{SSIM}$ (Wang et al., 2004) preserves 36/40 half-floor categories, a separate check outside the weight domain.

Connecting response to task consequences. Under the same 23 anchored permutations and original future targets, we compare native-to-perturbed response Q_{nat} with excess MSE ΔR. For $\delta = z^{\pi} - z$, the identity $\Delta R = \langle \delta^2 \rangle + 2\langle (z - y)\delta \rangle$ separates movement from alignment with native error. At mean ten, PredRNN exceeds ConvLSTM in both response (1.528 versus 0.565, $\times 10^{-3}$) and excess MSE (41.23 versus 4.75, $\times 10^{-6}$). Relative MSE increases are 3.05% versus 0.34%. This concordance holds in 4/5 subjects (subject 24 is the exception) and after every subject deletion; the other two model pairs fail the same symmetric stability screen (Appendix S15). A prospective test fixed this pair and direction before acquiring 720 forecasts on fifteen nonoverlapping later windows from the same videos. It meets the criterion: 5/5 subjects and all five deletion sets retain both positive differences. New-window relative MSE increases are 16.65% versus 0.049%, respectively. ConvLSTM's alignment offsets 98.7% of its squared movement, explaining why response magnitude alone does not determine excess error (Appendix S16). The earlier native-model selection gain still reverses after deleting one subject, and pooled task-error rankings remain unchanged (Appendices S13.6, S13.7).

5 Discussion

What the controls identify. The same shared reference weights reverse a predictor ordering across new segments, although central rules can agree. Thus reproducibility at a fixed rule and invariance to reference choice are distinct properties. The task analysis adds a complementary check: in the pooled mean-ten analysis under one fixed perturbation distribution, PredRNN's greater response than ConvLSTM accompanies a greater increase in target error, with prospective confirmation on later windows from the same videos. Together, these findings connect reference selection, repeatability and task consequences while keeping the two response definitions explicit.

Choosing the domain. Exact input-distance matching restricts eligible comparisons; endpoint balance cancels additive history effects; shared weights impose one comparison rule across clips. These requirements define the comparison problem whose weight domain is analyzed. The floor path then separates conclusions supported without a minimum edge mass from those requiring one. The domain describes weight choices within a fixed graph and readout, rather than an observed distribution of evaluator preferences. Changing the graph, readout or sharing convention asks a different question. The SSIM comparison, prospective delete-one-clip result and shared/product contrast test these distinct sources of variation separately (Appendix S14). Their agreement cannot substitute for weight invariance, and weight invariance cannot establish their agreement.

Relation to prior controls. Constrained surrogates and transition-preserving sequences (Schreiber & Schmitz, 2000; Pethel & Hahs, 2014), temporal-order verification (Misra et al., 2016), atemporal probes (Buch et al., 2022) and static-content controls (Liu et al., 2024) motivate specifying what an intervention holds fixed. Evaluation criteria can yield different conclusions about generative models (Theis et al., 2016). Our analysis concerns the reference comparisons left open after such input constraints. Sharing, weighting and readout choices determine different aspects of that comparison; Appendix S18 distinguishes this question from recognition and representation diagnostics.

6 Conclusion

Input controls can leave a predictor ordering dependent on the reference rule; a frozen witness pair reproduces this dependence on new segments. Under a separate fixed perturbation distribution, one response ordering aligns with excess forecast error and reproduces on prospectively specified later windows. Stable comparisons and agreeing central rules complete the account. Reporting the reference rule, its admissible range, local sensitivity and measured task consequence distinguishes repeatable probe results from broader claims about temporal ability.

AI use statement

We used OpenAI Codex to assist with research planning, experiment implementation and execution, numerical analysis, and manuscript drafting and revision. We used Claude for automated research and manuscript criticism. The experimental predictions were generated by the forecasting models

specified in this paper. Reported numerical results were checked against saved outputs using separate calculation and verification procedures, as detailed in the supplementary evidence. The authors take responsibility for the final content, including AI-assisted text, code, and claims.

Reproducibility statement

The supplementary evidence specifies the constructions, model checkpoints, preprocessing, reference constraints, and experimental comparisons. Appendices S15 and S16 detail the response–error decomposition and the later-window confirmation, including the prespecified criterion and complete subject and deletion results. Saved-output calculations, execution records, and independent numerical checks support the reported results.

References

Ali J Alrasheed, Aryan Yazdan Parast, Basim Azam, James Bailey, and Naveed Akhtar. Latent video prediction learns better world models. *arXiv preprint arXiv:2605.15618*, 2026. URL https://arxiv.org/abs/2605.15618.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://stanford.edu/~boyd/cvxbook/.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Fei-Fei Li, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2022. URL https://arxiv.org/abs/2206.01720.

Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. ContextVP: Fully context-aware video prediction. In *Proceedings of the European Conference on Computer Vision*, pp. 781–797, 2018. doi: 10.1007/978-3-030-01270-0_46. URL https://arxiv.org/abs/1710.08518.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in Fréchet Video Distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7277–7288, 2024. URL https://arxiv.org/abs/2404.12391.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://arxiv.org/abs/2311.17982.

Jaehui Hwang, Huan Zhang, Jun-Ho Choi, Cho-Jui Hsieh, and Jong-Seok Lee. Temporal shuffling for defending deep action recognition models against adversarial attacks. *Neural Networks*, 169: 388–397, 2024. doi: 10.1016/j.neunet.2023.10.033. URL https://arxiv.org/abs/2112.07921.

Siyuan Li, Cheng Tan, Zhangyang Gao, and Stan Z Li. OpenSTL: Open-source toolbox for spatiotemporal predictive learning. https://github.com/chengtan9907/OpenSTL/tree/OpenSTL, 2023. Original OpenSTL implementation.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1605.08104.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pp. 527–544, 2016. doi: 10.1007/978-3-319-46448-0_32. URL https://arxiv.org/abs/1603.08561.

Shawn D. Pethel and Daniel W. Hahs. Exact significance test for Markov order. *Physica D: Nonlinear Phenomena*, 269:42–47, 2014. doi: 10.1016/j.physd.2013.11.014. URL https://arxiv.org/abs/1302.1500.

Farrukh Rahman. What does a temporal benchmark score measure? Decomposing channel use in video VLM evaluation. *arXiv preprint arXiv:2607.12304*, 2026. URL https://arxiv.org/abs/2607.12304.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970. doi: 10.1515/9781400873173.

Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142:346–382, 2000. doi: 10.1016/S0167-2789(00)00043-9. URL https://arxiv.org/abs/chao-dyn/9909037.

Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pp. 32–36, 2004. doi: 10.1109/ICPR.2004.1334462. URL https://www.csc.kth.se/cvap/actions/.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 843–852. PMLR, 2015. URL https://proceedings.mlr.press/v37/srivastava15.html.

Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z. Li. OpenSTL: A comprehensive benchmark of spatio-temporal predictive learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69819–69831, 2023. doi: 10.52202/075280-3060. URL https://papers.nips.cc/paper/2023/hash/dcbff44d11130e75d09d3930411c23e1-Abstract-Datasets_and_Benchmarks.html.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.01844.

Karam Tomotaki-Dawoud, Anna Hilsmann, Peter Eisert, and Sebastian Bosse. Temporal preservation over processing: Diagnosing and designing spatiotemporal single-stage video detectors. *arXiv preprint arXiv:2606.31421*, 2026. URL https://arxiv.org/abs/2606.31421.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng, and Xiaokang Yang. Structure preserving video prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1460–1469, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Xu_Structure_Preserving_Video_CVPR_2018_paper.html.

S1 SUPPLEMENTARY EVIDENCE

The following sections retain the complete study reports and detailed analyses supporting the compact main text. Their initial, prospective and post-hoc scopes are unchanged. Results are not independent merely because they appear in separate sections. The companion files retain all prior execution and numerical evidence; the R20 saved-output entry point and endpoint certificates accompany the shared-weight comparison.

Table 3: Outcome map across the initial and follow-up studies for SimVP, unmasked observations and first-frame prediction disagreement. Ratios compare matched to reference means, not prediction accuracy. The abstract describes the initial study; the prospective follow-up tests the positive direction using fixed rules. The final post-hoc row changes matched and reference weights to achieve endpoint balance.

Study / family	Reference	Q ratio	Status
Initial / A	All unmatched	1.1466	Initial contrast
Initial / A	Exact input L1	1.0633	Post-hoc comparison
Follow-up / A	Exact input L1	0.9996	Positive criterion not met
Follow-up / B	Exact input L1	0.8978	Positive criterion not met
Follow-up / B	Balanced exact	0.9870	Post-hoc changed weights

S2 INPUT CONTROLS AND RESPONSE MEASUREMENTS

Each Moving MNIST clip supplies ten historical and ten future 64×64 grayscale frames (Srivastava et al., 2015). We use the released OpenSTL SimVP+gSTA-S and ConvLSTM artifacts (Tan et al., 2023; Li et al., 2023), with one fixed checkpoint each. The supplied ConvLSTM cell retains $h_{\text{new}} = o_t \odot g_t$; this is an artifact comparison rather than a controlled test of canonical architecture families. Arm V uses unmasked input. Arm M zeros columns 24–39 before constructing each history. Both arms predict the same unmasked future.

Five shared suffix frames (family A). Let $N = [0, 1, \ldots, 9]$ be native history. For every permutation (a, b, c) of $(0, 1, 2)$ define

$$P_{abc} = [0, a, b, c, 0, 5, 6, 7, 8, 9]. \tag{8}$$

The six histories share their frame multiset, first frame and final five native frames. At the index level their directed adjacent-pair multisets form three classes:

$$\{P_{012}, P_{120}\}, \quad \{P_{021}, P_{210}\}, \quad \{P_{102}, P_{201}\}.$$

For example, the last class contains the directed pairs $01, 10, 02, 20$, followed by the common pairs $05, 56, 67, 78, 89$. Every matched pair A, B therefore satisfies, for a position-independent function ϕ,

$$\sum_{t=0}^{8} \phi(x_t^A, x_{t+1}^A) = \sum_{t=0}^{8} \phi(x_t^B, x_{t+1}^B). \tag{9}$$

This is a direct multiset identity, not a new theorem. It covers any additive directed-pair summary, including signed and absolute pixel differences. Temporal assignment remains free. Histories omit native frames 3 and 4 and repeat frame 0, so the construction changes native motion and is not a natural-video intervention.

We enumerate all 15 unordered pairs, rather than selecting one unconstrained shuffle. Three pairs match directed edges at the index level and twelve do not. We classify the actual pixel sequences separately after masking; coincident frames can merge classes. All histories remain in the analysis. Consequently, a comparison within an edge class isolates a limitation of that input summary, while a comparison between classes provides a reference under the same frame content and suffix. These groups are not randomized treatments: their difference is not a causal effect of edge matching, and the unmatched reference is not an upper bound.

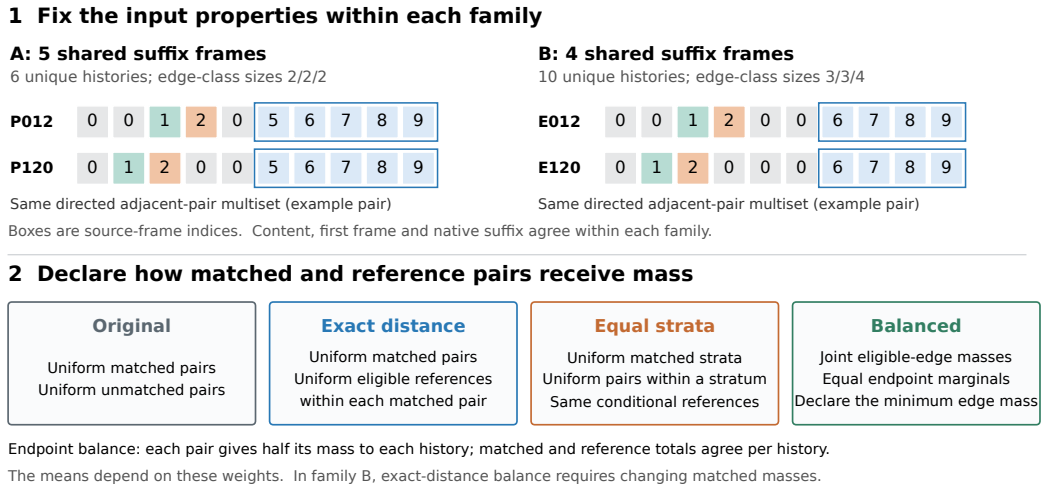

Figure 3: Input controls and comparison weights. Top: representative matched pairs; numbered boxes identify source frames, and outlines mark each family's common suffix. Content agrees within each family, not between families. The full families contain six and ten unique histories (family B is defined in Section S5). Bottom: four weighting definitions for fully covered graphs; strata index matched-pair distance coefficients. Initial-study uncovered pairs instead retain zero mass. Endpoint balance constrains each history's total mass, while the pairwise responses can still differ; Section S7 defines the balanced-weight domains.

Study grid. An earlier 32-clip study (indices 288–319) supplied 1,152 forecasts, and a planned exploratory short-block follow-up supplied 256 on those same clips. The latter compared P_{102} and P_{201} and motivated the present design. We then froze clips 320–351, both artifacts, both arms, native history and all six permutations, yielding $32 \times 2 \times 2 \times 7 = 896$ new study forecasts. The new native-history hashes do not overlap the 320 previously recorded study histories. All forecasts are saved before the new future targets enter scoring. Two separate synthetic smoke forecasts are engineering checks. No training or checkpoint selection occurs in this study. The benchmark test file served as validation upstream, so new study clips do not establish independent model selection.

Response at three aggregation levels. The initial study's primary error is raw full-image first-future MAE, $E(z, y) = \langle |z - y| \rangle$, using FP32 predictions and uint8 targets divided by 255 in float64 arithmetic. Equal-weight ten-frame averages and clipped predictions are secondary readouts. For a fixed pair, oriented from lexicographically earlier A to later B, define

$$d = E_B - E_A, \qquad D = |d|, \tag{10}$$

$$Q = \langle |z_B - z_A| \rangle, \qquad L = \langle \big| |z_B - y| - |z_A - y| \big| \rangle. \tag{11}$$

The triangle and reverse triangle inequalities give $D \leq L \leq Q$. Thus Q measures prediction disagreement, L retains changes in pixelwise error, and D retains only the difference of aggregate errors. Their differences locate losses from these readouts; they do not rank predictor quality. We report absolute prediction errors and ratios to same-arm native error alongside all responses. A deterministic predictor that uses only the suffix shared within a family produces identical forecasts for every history, hence $Q = L = D = 0$; this probe cannot distinguish such a degenerate predictor from any other suffix-only predictor.

For each clip, matched and unmatched pair responses are averaged separately, using their actual pixel-based membership; these clip means then receive equal weight. Every fixed named-pair summary includes all 32 clips. The two artifacts share clips, and the 15 pair responses share forecasts, so neither artifact count nor pair count enlarges the sample size. Means, medians, signs and complete leave-one-clip mean ranges describe this finite set without population confidence intervals or significance tests. A zero denominator produces an undefined ratio, never an added stabilizing constant.

Spatial and temporal accounting. For $c = \mathrm{clip}(z, 0, 1)$ and $y \in [0, 1]$, the pointwise identity $|z - y| = |c - y| + |z - c|$ separates bounded prediction error from output excursion. We further split clipped error over $y > 0$ and $y = 0$, keeping a full-image denominator for additive contributions and separate area-normalized densities. At every future step we save d_t, Q_t, L_t, D_t. Temporal cancellation compares $\mathrm{mean}_t\, |d_t|$ with $|\,\mathrm{mean}_t\, d_t|$, retaining all ten steps. These simple identities specify what is measured rather than claiming a temporal mechanism.

S3 An Initial Excess Under Complete References

All 896 initial-study predictions are finite and retained. Figure 4 shows the unmasked first-frame results; masked and ten-frame readouts remain separate comparisons. Fixed-pair error checks, including the higher native-relative error of the supplied ConvLSTM artifact, are in Appendix S19.1.

Matching does not guarantee a milder response. The complete unmasked SimVP grid has mean matched/unmatched $Q = 0.003891/0.003393$, $L = 0.003665/0.003203$ and $D = 0.002027/0.001750$. The corresponding ratios of means are 1.147, 1.144 and 1.158. The positive first-frame Q direction does not repeat under prospective exact-distance control (Table 3). Matched values exceed unmatched values on 27/32 clips for Q and L, and 22/32 for D. Every leave-one-clip mean difference remains positive: for Q, the range is 0.000443–0.000569. This is finite deletion sensitivity, not a confidence interval. Across both arms and both horizons, SimVP's twelve group-level response ratios range from 1.071 to 1.205 and all exceed one (Appendix S19).

ConvLSTM's unmasked first-frame matched/unmatched ratios are 0.998, 0.996 and 0.970 for Q, L, D. Its Q means differ by only -0.000237, with 15 positive and 17 negative per-clip group differences. Aggregate-error ratios vary more across the grid, from 0.833 to 0.989, while output-disagreement ratios stay between 0.997 and 1.001. The near-equality of a group mean is a measured comparison, not an equivalence test. The complete 15-pair tables prevent a favorable pair from standing in for the grid.

For the ordinary three two-member classes, each history occurs once among matched pairs and four times among unmatched pairs. Consequently, the two groups' average pair error is exactly the same six-history average by construction. Their response differences cannot be explained by a different weighting of high-error endpoints. They can still reflect other differences between pairs, including changed temporal assignment; the comparison does not isolate a causal matching effect.

Actual pixels can merge nominal classes. Input checks identified one degeneration before inference: clip324 in arm M has classes of sizes four and two, giving seven matched and eight unmatched pairs. The histories P_{102} and P_{201} become identical, and both models produce exactly zero response for that pair at both horizons. This observation remains in every relevant summary. Other 63 clip/arm combinations retain the three ordinary classes. We use actual pixel equivalence for group membership; the balanced endpoint argument above is not applied to the exceptional masked case.

Aggregation hides different amounts of response. For the fixed unmasked pair P_{102}, P_{201}, mean D/Q measured as a ratio of means is 0.542 for SimVP and 0.188 for ConvLSTM. Prediction differences can therefore be much larger than the differences retained by the aggregate error. Temporal aggregation adds another loss: comparing $\mathrm{mean}_t\, |d_t|$ with $|\,\mathrm{mean}_t\, d_t|$ removes 18.1% of SimVP's and 3.51% of ConvLSTM's mean stepwise response in arm V. The corresponding M values are 5.44% and 12.95%. All ten steps and all fifteen pairs are supplied; no future step is chosen to maximize an effect. Earlier saved-output analyses additionally localize clipped spatial contributions and output excursion (Appendix S27.1).

S4 Distance Matching Weakens the Initial Contrast

The following analyses were specified after the first review of the completed grid. They reuse its saved inputs and outputs with zero new forecasts, and remain descriptive post-hoc checks.

Exactly equal input distance. The original groups differ in within-pair perturbation size. For each matched pair, we therefore average all unmatched pairs in the same clip and arm with the exactly

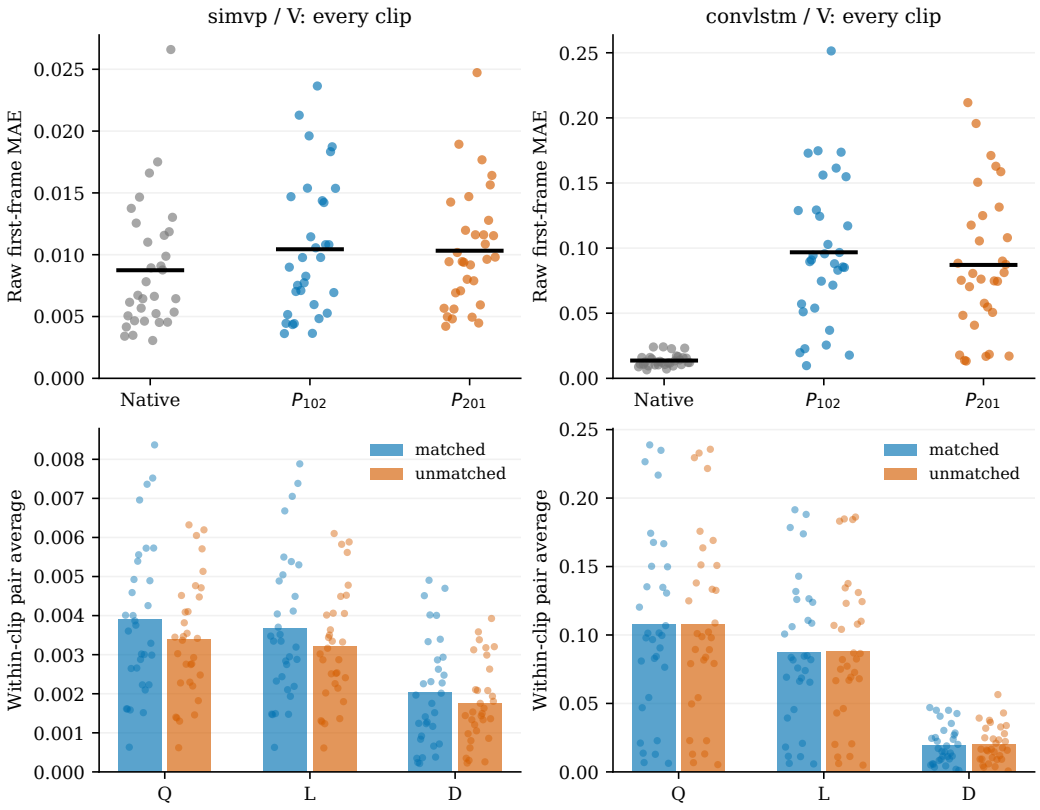

Figure 4: Initial study (clips320–351), arm V. Top: raw first-frame errors for native history and the prespecified matched pair; black marks are means and points retain every clip. Bottom: matched/unmatched within-clip pair averages for Q, L, D; bars are means and points retain all 32 clip averages. Models and rows use different vertical scales, each starting at zero. These are response magnitudes, not performance improvements.

equal integer input L1 sum. A stricter rule also requires equality at every pixel of the ten-frame absolute-difference-sum map. We average references within a matched pair, covered matched pairs within a clip, then clips equally. Every unmasked matched comparison has references: 96/96 across all 32 clips.

SimVP's unmasked first-frame Q, L, D ratios become 1.063, 1.064 and 1.079 under scalar matching, compared with 1.147, 1.144 and 1.158 for the original groups. The Q difference is positive on 23/32 clips. Pixelwise-map matching gives a Q ratio of 1.061. For ConvLSTM, the scalar-matched ratios are 0.992, 0.991 and 0.944; its pixelwise-map Q ratio is 0.994. Input-distance control thus reduces the SimVP contrast without reversing its unmasked mean direction in either rule. This does not isolate an effect of transition matching: temporal placement and other pair properties remain different.

The masked result is less consistent. Scalar-matched SimVP first-frame ratios are 0.997, 0.999 and 1.000 (the last is rounded from below one). Three zero-distance matched pairs in clip324/M have no unmatched zero-distance reference. We report 97/100 covered matched comparisons, without forcing references or treating this subset as the original group mean. Both rules, all readouts and full coverage records appear in Appendix S20.

The initial positive Q comparison is not repeated by the prospective exact-distance test in Table 3.

Output range and future steps. For interval clipping $c(z)$, the signed residual $e(z) = z - c(z)$ and $c(z)$ are both nondecreasing. Hence $|z_B - z_A| = |c_B - c_A| + |e_B - e_A|$ pointwise, yielding $Q = Q_c + Q_e$. In the unmasked first-frame matched groups, Q_e/Q is 1.76% for SimVP and 25.92% for ConvLSTM. Clipping leaves matched/unmatched Q_c ratios of 1.147 and 0.998: ConvLSTM's

near-flat group comparison is therefore also present inside the valid output range, rather than being explained solely by excursion. This is output accounting, not evidence for internal saturation. Across all ten individual steps of the original unmasked grid, SimVP's Q, L, D ratios exceed one; masked Q, L reverse at step three. These stepwise comparisons have not been distance-reweighted. Complete curves and range accounting are supplied in Appendix S20.

S5 Prospective Controls Do Not Confirm the Initial Excess

The initial study above and in the abstract comprises 896 forecasts on clips 320–351. We next fixed a separate study slice, clips 352–383, before obtaining any outputs or future targets. The two existing checkpoints and V/M arms were retained. This follow-up jointly specified both exact-distance rules, first-frame Q as the primary readout, all ten future steps, clipped outputs, full coverage and the same reference weights throughout. It asks whether the initial SimVP/V first-frame excess persists, rather than selecting a favorable secondary readout afterwards.

Four shared suffix frames (family B). Family A is the six-history construction in Eq. 8. Family B enumerates, for the same six permutations,

$$E_{abc} = [0, a, b, c, 0, 0, 6, 7, 8, 9], \tag{12}$$

$$L_{abc} = [0, 0, a, b, c, 0, 6, 7, 8, 9]. \tag{13}$$

The twelve names give ten unique histories because $E_{012} = L_{120}$ and $E_{021} = L_{210}$. Their directed-edge classes have sizes 3, 3 and 4, yielding twelve matched and thirty-three unmatched pairs. Each same-name E/L pair preserves directed edges. Histories share their content, first frame and final four native observations within B. We forecast each unique input once and retain aliases only as explicit mappings. Native history is shared across A/B, giving $32 \times 2 \times 2 \times 17 = 2{,}176$ study forecasts, plus two separate engineering checks, with zero training.

Every new matched pair has an exact-distance reference: 96/96 for A and 384/384 for B in each arm. The scalar and pixelwise-map rules select exactly the same reference graph on this slice. They therefore constitute one realized comparison with two satisfied constraints, not independent confirmations. All 32 histories differ from the 352 previously recorded native histories by exact hash; digit identities and upstream model-selection dependence are not thereby separated.

Table 4: Prospective slice, first-frame prediction disagreement. Original denotes all unmatched pairs; exact distance uses the frozen hierarchical reference weights. Both exact-distance rules yield identical reference sets on this slice. Ratios below one indicate smaller matched disagreement, not better prediction. Signs count positive/negative per-clip exact-distance differences. Coverage is 96/96 for A and 384/384 for B in each arm, across all 32 clips.

Artifact	Arm	Family	Original Q	Exact Q	Exact, clipped Q	$+/-$
SimVP	V	A	1.0604	0.9996	0.9993	14/18
SimVP	V	B	1.0659	0.8978	0.8978	8/24
SimVP	M	A	1.0735	1.0185	1.0182	17/15
SimVP	M	B	1.0493	0.8593	0.8580	4/28
ConvLSTM	V	A	1.0027	1.0006	1.0012	18/14
ConvLSTM	V	B	1.0206	1.0047	1.0061	21/11
ConvLSTM	M	A	1.0113	1.0108	1.0126	17/15
ConvLSTM	M	B	1.0268	1.0224	1.0291	14/18

The positive primary direction does not repeat. For SimVP/V/A, the original-group Q ratio is 1.0604. Exact-distance references give matched $Q = 0.00389768$ and reference $Q = 0.00389919$, a ratio of 0.9996. The clip differences have 14 positive and 18 negative signs; deleting one clip gives ratios 0.9813–1.0125. The prespecified criterion required a ratio above one under both rules with complete coverage. It is not met. This near-parity mean is not an equivalence test. The secondary mean10 ratio is 1.0209; it does not rescue the failed first-frame criterion. ConvLSTM/V/A is also near parity at the primary horizon, with ratio 1.0006.

The second construction changes the ordering. The reversal in B is concentrated in coefficient stratum $(2, 2, 0)$, whose ratio is 0.8118; stratum $(1, 1, 1)$ remains above one at 1.0265. Under the fixed weights, the overall SimVP/V/B ratio changes from 1.0659 with all unmatched references to 0.8978 with exact-distance references: matched $Q = 0.00452860$ versus reference $Q = 0.00504387$. Differences are negative on 24/32 clips (Figure 5), and all delete-one ratios remain below one, 0.8835–0.9076. Its masked ratio is 0.8593, and the exact-distance Q ratio remains below one at each of the ten steps in both arms. These are finite comparisons within B; the result is not a uniform layout effect or a causal A-to-B improvement, since frame multiplicity and the common suffix change together.

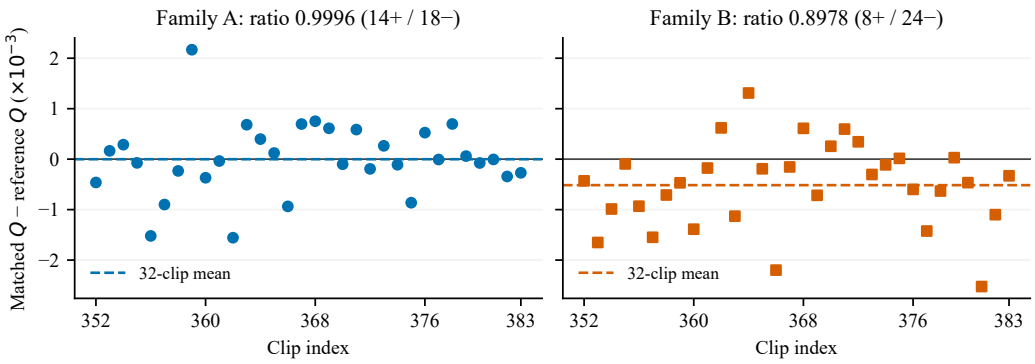

Figure 5: All 32 prospective clip differences for the primary SimVP/V first-frame readout, using exact-distance references. Points retain clip order and both panels share the same scale; dashed lines mark means, not confidence bounds. The ratio in each title is a ratio of means. Negative differences indicate smaller matched disagreement, not better prediction. Both exact-distance rules select the same references.

Reference weights and readouts matter. B does not inherit A's original endpoint-balance identity. Equal mass is assigned to each matched pair, then split equally over that pair's references. Six of B's twelve matched pairs lie in stratum $(2, 2, 0)$ and each uses the same two references: together those references receive $6/12 = 1/2$ of the total reference mass. The remaining stratum $(0, 0, 2)$ has ratio 0.9687. We report every stratum and all six E/L contrasts in Appendix S21. Replacing the full unmatched mean with an exact-distance reference changes pair and endpoint weights, so the shift cannot be attributed to distance alone.

Using the same distance weights after clipping gives SimVP/V ratios 0.9993 for A and 0.8978 for B. The reversal is therefore present in bounded-output comparisons too. Readouts still differ: for A, Q is near parity while D has ratio 1.0339. The joint ten-step and range accounting prevents a secondary scalar-error pattern from being reported as persistence of the primary prediction-disagreement result.

S6 When Exact-Distance References Cannot Be Balanced

We next analyze the saved follow-up outputs without additional forecasts. These post-hoc analyses separate reference weighting from construction choice; they do not revise the prospective criterion. The weight schemes and complete subfamily enumeration are defined by input properties and applied to previously examined outputs.

The reference shift is between distance-coefficient strata. Write a pair's input-distance map as $n_{01}|F_0 - F_1| + n_{02}|F_0 - F_2| + n_{12}|F_1 - F_2|$. On every follow-up clip and arm, the exact reference graph equals the graph defined by equal coefficient triples. References remain uniform within each triple. Consequently, the change from the original unmatched mean to the exact-distance mean is entirely a redistribution of mass between coefficient strata; no within-stratum change remains. For SimVP/V/first, the reference Q rises from 0.00367582 to 0.00389919 in A and from 0.00424874 to 0.00504387 in B. The matched means remain fixed. This is an exact accounting of the comparison, not an isolated effect of distance (Appendix S22).

Endpoint balance has a feasibility condition. Assign half of each pair's mass to each of its two histories. Equal endpoint marginals make every weighted pair-average of a single-history score agree between matched and reference groups, while leaving pairwise responses such as Q unconstrained. A retains this balance under its exact references. In B, the four-history edge class receives half of the matched endpoint mass. On eligible exact-distance support, only coefficient $(1, 1, 1)$ references touch this class, at one endpoint; the four such matched pairs force reference endpoint mass $\frac{1}{2} \cdot \frac{4}{12} = \frac{1}{6}$. Thus no choice of conditional reference weights can balance B while preserving equal matched-pair mass and this exact support. The obstruction is visible in the inputs before forecasting.

A general input-only test. Let a_s be the two-endpoint incidence vector of pair s, and let R_m be the eligible references for matched pair m. For fixed matched masses u_m, a positive-mass row with no reference is infeasible. Otherwise, conditional reference weights can balance the histories exactly when, for every history potential f,

$$f^T \sum_m u_m a_m \;\le\; \sum_{m:u_m>0} u_m \max_{r \in R_m} f^T a_r.$$

A violating potential is an input-only certificate that no conditional reweighting can preserve this matched comparison and balance its endpoints. This follows from convex separation, not a new optimization theorem (Boyd & Vandenberghe, 2004). If both sides may change, feasibility is instead $0 \in \mathrm{conv}\{a_m - a_r : (m, r) \in E\}$; keeping every eligible joint edge positive requires relative-interior membership (Rockafellar, 1970). The complete statements and proofs are in Appendix S24.

With exclusively cross-class references and every such reference available to every row, the fixed-mass criterion reduces to a maximum class endpoint mass of one half. Under exact-distance restrictions this is only necessary. All 54 six-history subfamilies pass this class test, yet only eight admit balance with uniform matched mass. Fourteen admit flexible balance, but only eight can retain positive mass on every matched pair, and only six can retain every eligible joint edge. Thus feasibility, matched coverage and joint-edge coverage are distinct questions, even with the same class sizes.

Table 5: Input-graph feasibility counts. Fixed preserves each graph's recorded matched masses; flexible allows both matched and reference masses to change. Positive-M and positive-E are separate flexible-weight requirements on every matched pair and every eligible joint edge, respectively. These are reused input graphs, not independent prediction samples.

Input set	N	Existing	Fixed	Flexible	Positive-M	Positive-E
B subfamilies	54	6	8	14	8	6
Initial: original	64	63	63	64	64	64
Initial: scalar	64	62	63	63	63	63
Initial: map	64	63	63	63	63	63

The initial study illustrates three different obstructions. Scalar matching on unmasked clip 340 can be balanced without changing matched masses. The original comparison on masked clip 324 cannot preserve its matched masses, but flexible weights can balance it. Neither exact-distance graph on masked clip 324 permits any normalized balanced weights, even allowing zeros. Across all 192 initial graphs, 188 existing comparisons are balanced, 189 can be balanced while preserving their recorded matched masses, and 190 admit flexible balance. The feasibility checks do not replace the original ratios or add forecasts; they identify which comparison definitions can be retained.

A feasible comparison changes the weights being compared. We compare four schemes: *original* gives uniform mass to all matched and all unmatched pairs; *exact* keeps matched pairs uniform and splits each one's mass equally over its distance-matched references; *equal strata* first divides matched mass equally among its coefficient strata, then equally among pairs within each stratum, retaining the same conditional references; *balanced* uses the joint witness below. The witness maximizes the minimum joint mass on all eligible matched-reference edges, using only the symbolic input graph; it keeps every matched pair and eligible reference edge positive. In B it necessarily changes matched as well as reference weights. For SimVP/V/first, the resulting Q ratio is 0.9870, compared with 0.8978 under the original exact-distance weights; the masked ratios are 0.9726 and 0.8593. The selected witness need not be the unique optimum, so these are sensitivity comparisons, not robustness over all

balanced weights or evidence that balancing causes the difference. Weight dependence also affects the near-parity ConvLSTM comparisons: in A, equal-stratum first-frame Q ratios are 1.0520 (V) and 1.0653 (M), versus 1.0006 and 1.0108 under exact weights. Thus the interpretation of near-parity is specific to a weight scheme.

Table 6: Post-hoc first-frame Q ratios on the saved follow-up outputs. Original and exact retain equal matched-pair mass; equal strata change both matched and reference weights. The balanced witness changes both in B and coincides with exact weights in A. Balanced denotes the fixed input-only max-min support witness, not all possible balanced weights. All rows retain 32 clips.

Artifact	Arm	Family	Original	Exact	Equal strata	Balanced
SimVP	V	A	1.0604	0.9996	0.9891	0.9996
SimVP	V	B	1.0659	0.8978	0.9249	0.9870
SimVP	M	A	1.0735	1.0185	1.0068	1.0185
SimVP	M	B	1.0493	0.8593	0.8766	0.9726
ConvLSTM	V	A	1.0027	1.0006	1.0520	1.0006
ConvLSTM	V	B	1.0206	1.0047	1.0171	0.9964
ConvLSTM	M	A	1.0113	1.0108	1.0653	1.0108
ConvLSTM	M	B	1.0268	1.0224	1.0403	1.0007

All two-per-class subfamilies, with content and suffix fixed. Choosing two histories from each of B's three edge classes gives $\binom{3}{2}\binom{3}{2}\binom{4}{2} = 54$ six-history subfamilies. All preserve B's content multiset and suffix. Forty-six cover every matched pair with exact references; six also retain endpoint balance under uniform conditional references. We evaluate all 46, including the 40 unbalanced ones, with all fixed readouts. For SimVP/V/first, ratios among the six balanced subfamilies range from 0.9477 to 1.0276, with four above one and two below. Across all 46, the range is 0.8033–1.0596, with nine above one. Endpoint balance therefore does not impose a response ordering, and these mixed directions are not a positive confirmation. The subfamilies overlap and reuse the same 32 clips and predictions.

S7 Does Balance Determine the Comparison?

A single feasible weighting does not establish stability over the other balanced choices. We therefore compute the complete range on the saved follow-up outputs. Let $\bar{z}_m, \bar{z}_r$ be a pair's readout averaged over the same 32 clips, and let a_m, a_r denote its history-incidence vectors. On the exact-distance support E, define

$$\mathcal{P}_\tau = \left\{ x : \sum_{(m,r)\in E} x_{mr} = 1, \quad \sum_{(m,r)\in E} x_{mr}(a_m - a_r) = 0, \quad x_{mr} \geq \tau \right\}.$$

For $x \in \mathcal{P}_\tau$, the comparison is $\rho_z(x) = \sum x_{mr}\bar{z}_m / \sum x_{mr}\bar{z}_r$. The same x applies to all clips; it cannot choose a different comparison for each clip. We minimize and maximize ρ_z separately for each fixed artifact, arm, family, horizon and readout. These outcome-dependent extrema describe sensitivity; they are not newly selected input-only weights or additional confirmation results.

We fix three domains before computing the ranges: the nonnegative closure $\mathcal{P}_0$, a positive floor $\mathcal{P}_{t_\star/2}$, and the entire maximum-minimum-mass face $\mathcal{P}_{t_\star}$. Here $t_\star = 1/12$ for A and $1/88$ for B (Appendix S22); the half-floor is an explicitly chosen sensitivity setting, not a calibrated recommendation. Exact rational certificates verify every endpoint and strictly positive reference-mean denominators on all domains. The arithmetic is exact for the archived decimal scalars. No sampling uncertainty is represented by these intervals.

The complete grid separates stable and weight-dependent directions. Table 7 summarizes all ranges. In the nonnegative closure, 111 of A's 288 cells and 278 of B's 288 cells cross one; the others remain wholly on one side. On the maximum-minimum face, all A cells are one-sided, whereas 25 B cells still cross. Thus neither universal instability nor universal direction stability follows from endpoint balance. These counts summarize correlated readouts, not a population prevalence.

Table 7: Direction of the complete weight-sensitivity ranges. Each row covers 288 correlated artifact/arm/horizon/readout cells in one family; first and step01 duplicate a readout. Counts use the prespecified 1 ± 10^{-10} display band. No cell touches or stays within that band. These are counts within the declared grid, not frequencies of independent trials.

Family	Minimum edge mass	Wholly below	Crosses both sides	Wholly above
A	0	114	111	63
A	$t_\star/2$	137	43	108
A	$t_\star$	155	0	133
B	0	10	278	0
B	$t_\star/2$	52	236	0
B	$t_\star$	238	25	25

Balance permits opposing directions. For SimVP under unmasked observations, first-frame raw Q has closure ranges of 0.9653–1.0365 in A and 0.9057–1.0614 in B. Boundary witnesses can assign zero weight to eligible edges. Requiring every edge to have mass at least $t_\star/2$ narrows these to 0.9822–1.0177 and 0.9416–1.0229, respectively; both still cross one. Thus opposite response orderings are feasible with the same histories, exact-distance support, positive edge support, balanced endpoints and shared clip weights. Pairwise response is not fixed by the single-history marginal constraint.

The weighting objective narrows the range. On the full maximum-minimum face, A's primary ratio is fixed at 0.9996, while B ranges from 0.9843 to 0.9870. B's previous witness lies at the upper endpoint for this readout. The prospective equal-matched ratio 0.8978 is outside these balanced domains because its weights are not endpoint-balanced; it remains a valid fixed-protocol result. The envelopes examine a changed estimand, rather than invalidating that protocol's observed ordering. B's balanced ratio remains below one throughout the maximum-minimum face, but not over all balanced positive weights. This distinction also matters for ConvLSTM/M/first: B's face ranges from 0.9960 to 1.0007 and still crosses one. Figure 6 retains both artifacts and observation arms. The complete grid of 1,728 envelope rows (576 settings across three domains) includes first, mean10 and all ten steps, raw/clipped Q, L, D, and all three domains; these rows are correlated analyses of the same saved forecasts.

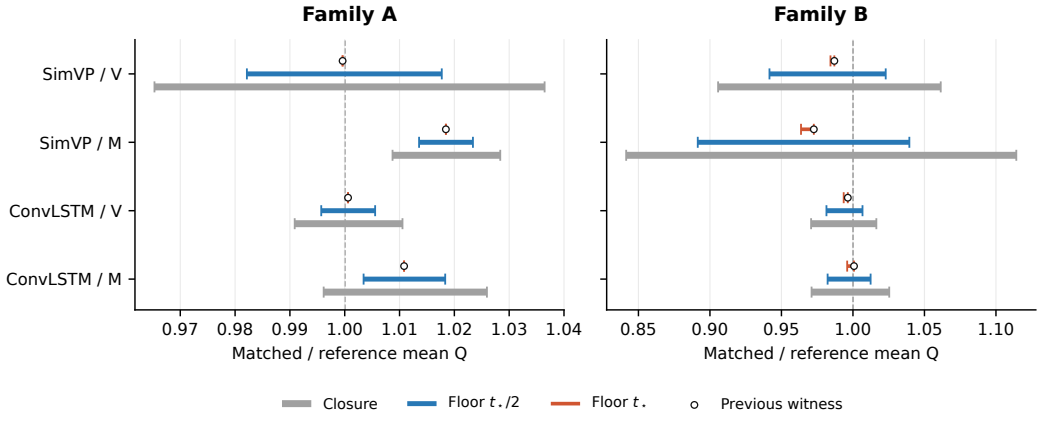

Figure 6: Complete first-frame raw Q ranges under three balanced-weight domains; circles mark the previous input-defined witness. Gray bounds may require zero edge mass; blue and orange domains retain every eligible edge. Horizontal scales differ between families. Ranges are optimization bounds over one finite set of outputs, not uncertainty intervals or model-quality rankings.

Realized graphs must also be checked in the initial study. The follow-up's scalar, spatial-map and coefficient supports agree on all 128 input graphs. The initial study has two exceptions: unmasked clip 340 gains three extra scalar-distance reference edges and loses endpoint balance, while stricter

map matching retains the coefficient graph and balance. Masked clip 324 has merged pixel classes and partial matched-pair coverage; both exact rules remain unbalanced. Across the initial 32 clips, scalar matching is balanced on 31/32 in each arm, whereas map matching is balanced on 32/32 unmasked and 31/32 masked. These graph checks retain every clip and the original covered-pair weighting; no forecasts or ratios are recomputed or replaced (Appendix S23).

S8 Cross-Architecture Evaluation on Human-Action Video

We extend the same A/B constructions to KTH human-action video (Schüldt et al., 2004) with the official OpenSTL KTH20 SimVP+gSTA and ConvLSTM checkpoints (Tan et al., 2023). SimVP was evaluated first. After that result, ConvLSTM was specified as the second architecture, using byte-identical inputs, comparison graphs and weights. Thus this is a sequential cross-architecture test, not a jointly blinded two-model experiment. We perform no training or selection among checkpoints using their KTH probe outputs.

Fixed inputs and conditional support. The cohort comprises prediction-test subjects 17, 19, 21 and 23, all six actions, scenario d1, and the first 20 frames of each video's first annotated subsequence. Ten grayscale history frames at 128×128 produce ten retained future predictions. Six of the 24 realized graphs lack an unmatched reference group; both models retain these failures and use the identical input-defined 18-clip conditional comparison set. The complete 24-clip contrast is undefined. The earlier uniform correction from a 30-frame extraction requirement to 20 occurred before decoding and SimVP inference, with no video substitution. Actual-history deduplication yields 334 scientific calls per architecture, plus one synthetic engineering call each. The ConvLSTM round reuses all SimVP outputs without replay. Total KTH scientific calls are 668, with zero training.

Domain and readout. The primary readout is first-future raw Q, with equal weights over the 18 supported clips. Each graph's eligible-edge mass is at least $t_*/2$, with endpoint balance. The default and input-frozen max-min witness are distinct estimands; neither is a model-quality score. B's default weights are outside the balanced domain. Each clip independently chooses its admissible weights, so the aggregate signed-difference interval is the exact average of the clipwise minima and maxima. This product domain differs from the single shared weight vector used for Moving MNIST in Section S7; widths cannot be directly compared across those domains. All bounds concern fixed outputs, not sampling uncertainty.

Table 8: KTH results on the identical 18/24 conditional subset. Ratios use clip-averaged Q. The last column is a signed $Q_M - Q_R$ range, scaled by 10^5, not a ratio interval. First is primary; mean 10 is secondary. Endpoints are exact relative to archived binary64 pair costs recomputed from saved float32 outputs.

Model	Family	Horizon	Default	Witness	Difference range $\times 10^5$
SimVP	A	first	1.1444	1.1444	[6.729, 8.221]
SimVP	A	mean 10	1.0992	1.0992	[7.364, 9.287]
SimVP	B	first	0.9225	1.0104	[-4.107, 6.579]
SimVP	B	mean 10	0.9448	1.0167	[-4.656, 9.837]
ConvLSTM	A	first	1.0379	1.0379	[0.023, 1.671]
ConvLSTM	A	mean 10	0.9592	0.9592	[-2.284, -0.239]
ConvLSTM	B	first	0.9151	0.9824	[-4.022, 2.155]
ConvLSTM	B	mean 10	0.8349	0.9034	[-10.579, 1.109]

What transfers, and what does not. Both predeclared primary domain propositions hold on the two KTH artifacts: A's full balanced interval is strictly positive, while B's interval contains both positive and negative contrasts. For ConvLSTM, A's default and witness ratio is 1.0379, compared with SimVP's 1.1444. ConvLSTM's default A and B contrasts are respectively $+3.8\%$ and -8.5% of their reference means. ConvLSTM's A lower bound is only 2.329×10^{-7} in raw Q units: strict positivity is an exact statement about this finite weight domain, not a large practical margin or robustness to new samples.

The specific default-to-witness reversal does *not* transfer. ConvLSTM B changes from 0.9151 to 0.9824, remaining below one at both points, whereas SimVP changes from 0.9225 to 1.0104. Nevertheless ConvLSTM B's full balanced range is $[-4.022, 2.155] \times 10^{-5}$. Thus failure of one predetermined witness to reverse the ordering does not establish stability over the full domain. Conversely, neither the witness reversal nor the optimized opposite direction identifies a scientifically correct weighting.

Weight stability does not imply horizon stability. The complete secondary results provide a second boundary. ConvLSTM A's mean-ten ratio is 0.9592 and its *entire* balanced difference interval is negative, $[-2.284, -0.239] \times 10^{-5}$, although its first-frame interval is entirely positive. SimVP A remains positive at both reported horizons. B spans both signs at both horizons in both models. The cross-architecture agreement is therefore specifically about the predeclared first-frame distinction, not a horizon-independent A advantage. All four subject-level A first-frame point contrasts are positive for both models, but this does not establish stability over weights within each subject; the bounds reported here aggregate the conditional 18 clips.

Table 9: First-future raw Q across two datasets and architectures; Moving MNIST uses unmasked inputs. Both use a half-floor, but the weight domains differ. Cells are signed-difference ranges scaled by 10^5. This is a finite comparison map, not a prevalence or performance ranking.

Dataset	Model	Clips	Domain	A range $\times 10^5$	B range $\times 10^5$
KTH	SimVP	18/24	clipwise	[6.729, 8.221]	[-4.107, 6.579]
KTH	ConvLSTM	18/24	clipwise	[0.023, 1.671]	[-4.022, 2.155]
Moving MNIST	SimVP	32/32	shared	[-7.083, 6.781]	[-29.222, 9.966]
Moving MNIST	ConvLSTM	32/32	shared	[-43.881, 56.300]	[-220.684, 79.429]

Under their respective stated domains, all four Moving MNIST model/family cells in Table 9 span both signs, whereas the two KTH A cells are one-sided. This compares sign patterns, not widths, and does not isolate whether dataset, checkpoint or domain differences account for the contrast.

Post-hoc horizon and subject sensitivity. Using only saved outputs, a complete ten-step A analysis finds ConvLSTM positive at steps 1–2, sign-indefinite at 3–4 and negative at 5–10 over the same declared domains. Its step-2 lower endpoint is merely 7.65×10^{-9}. SimVP is negative at step 3 despite being positive at every other step and in its mean-ten readout. These are horizon profiles, not evidence of a causal internal mechanism. A complete delete-one-subject check also shows that ConvLSTM A's first-frame range spans both signs after omitting subject 17 or 21, and its mean-ten range becomes positive after omitting subject 23. The original 18-clip results remain unchanged, but neither composition robustness nor universal horizon stability follows. Both post-hoc grids, including every omission and step, are retained in Appendix S26; they are not new prospective confirmations.

The extension tests direction stability under declared weights on two separately trained KTH artifacts. It does not establish improved forecasting, a causal order mechanism or downstream decision utility. The four subjects provide limited finite evidence, with unequal scorable counts; reordered histories remain artificial even though the source videos are real.

S9 Sharing Weights Across Clips

The KTH ranges above allow each clip to use its own balanced weights. We now ask whether the observed direction ambiguity requires this freedom. This is a post hoc reanalysis of the same saved outputs, with no new predictions or training. An input-only check establishes that all 18 supported clips have identical named edges and balance equations within each family: A has 10 edges with $t_* = 1/12$, and B has 66 with $t_* = 1/88$. We impose a single vector of named-edge weights across all clips, retaining the same half-floor. The six unsupported clips remain excluded; the full-24 contrast remains undefined.

Let D_f denote the per-clip feasible weight domain and c_i the signed edge costs for clip i. The new bounds optimize $\bar{c}^\top w$ over $w \in D_f$, where $\bar{c} = \sum_i c_i/18$. The earlier bounds optimize $\sum_i c_i^\top w_i/18$ over D_f^{18}. The diagonal embedding of the shared domain into the product domain implies containment

of the new interval. For mean ten, costs are time-averaged before optimization: one weight vector is shared across both clips and prediction steps. Exact endpoint certificates refer to the archived binary64 costs, not population uncertainty.

Table 10: Same saved KTH outputs under two weight domains, all eight cells. Signed $Q_M - Q_R$ ranges are scaled by 10^5. Product allows clip-dependent weights; shared requires one vector across all 18 clips. A narrower range changes the reporting conclusion, not forecast accuracy.

Model	Family	Horizon	Product domain	Shared domain
SimVP	A	first	[6.729, 8.221]	[6.764, 8.186]
SimVP	A	mean 10	[7.364, 9.287]	[7.769, 8.882]
SimVP	B	first	[-4.107, 6.579]	[-3.996, 6.512]
SimVP	B	mean 10	[-4.656, 9.837]	[-4.078, 9.158]
ConvLSTM	A	first	[0.023, 1.671]	[0.023, 1.671]
ConvLSTM	A	mean 10	[-2.284, -0.239]	[-2.267, -0.256]
ConvLSTM	B	first	[-4.022, 2.155]	[-3.848, 2.083]
ConvLSTM	B	mean 10	[-10.579, 1.109]	[-9.017, -0.313]

Which ambiguity requires clip-dependent weights? At the first-future readout, B still admits both signs with shared weights for both artifacts: SimVP gives $[-3.996, 6.512] \times 10^{-5}$ and ConvLSTM gives $[-3.848, 2.083] \times 10^{-5}$. Thus the primary B sign ambiguity does not require selecting weights separately for each clip. A remains positive for both artifacts, which follows from the already positive product-domain bounds and is not an independent replication.

For ConvLSTM B at mean ten, the product interval $[-10.579, 1.109] \times 10^{-5}$ spans zero, but its shared interval $[-9.017, -0.313] \times 10^{-5}$ is entirely negative. On these outputs, obtaining a positive mean-ten B contrast therefore requires relaxing the common-weight constraint within the declared product domain. SimVP B remains two-sided at mean ten. ConvLSTM A remains negative at mean ten, retaining its horizon-dependent sign change. All eight cells are reported, including this loss of sign ambiguity.

Evaluation consequence and limit. The exact reporting conclusion for ConvLSTM B mean ten changes from direction-ambiguous to uniformly negative when the probe must use one common weighting rule. Conversely, first-frame B ambiguity survives that restriction. This isolates which finite conclusions depend on clipwise adaptation, rather than attributing every sign change to it. Optimized weights are descriptive witnesses, not policies validated on fresh data. No result selects a scientifically correct weighting, establishes downstream utility, or compares model quality. Sharing weights aligns this convention with the Moving MNIST analysis, but does not make different datasets, graphs and checkpoints a controlled dataset comparison.

S10 Prospective New-Subject Reporting Validation

The design, global margin threshold, input admission and both-direction partial-row rule were fixed before 838 scientific forwards on five new study subjects. Two separate synthetic checks are excluded from that count. Both official artifacts completed 419/419 science calls, with no training, retries or future-target access. All 30 input clips are retained in the admission denominator; eight lack positive support and one has a different named comparison graph. Each construction therefore admits 21/30 clips. The same admission is used for all policies, artifacts and readouts.

The new subjects were absent from the earlier development cohort, but are part of the same upstream KTH prediction-test split. This does not establish independence from the original checkpoint authors' selection procedures. The first-frame readout is primary; mean ten is secondary. Subject-level cells share a person and cannot be treated as independent observations. Pooled rows weight eligible clips equally and are not added to the 40 subject-level cells when counting evidence.

The development threshold is exactly $3914016276/98343223193$. A report requires a strictly larger margin. The 32 development cells contain 16 unsupported balanced point reports; thresholding retains 14 reports without overstatement. No new-subject result changes this threshold. Full-domain

endpoints and all intermediate reduced-cost certificates are checked using exact rational arithmetic relative to stored binary64 costs.

In the tables, + and − denote a reported positive or negative direction; ? denotes abstention or a non-one-sided full interval. Legacy is the original reference weighting and need not target the balanced domain. Balanced is one input-defined feasible witness. U in the main text means that generalizing a point direction to the full domain is unjustified; a point contrast is not numerically wrong at its own weights. All new unsupported reports lie on two-sided intervals, rather than opposing a robust direction.

Table 11: New-subject validation, 20 correlated cells per readout. U: reports without full-domain direction support; Miss: robust directions withheld. The full-domain row defines support and its zero U is not a measured accuracy improvement. The fixed-row certificate has the same counts; its limited B coverage is discussed below.

Readout	Reporting rule	Reports	U	Miss
First (primary)	Legacy point	20	10	0
	Balanced point	20	10	0
	Calibrated margin	10	0	0
	Full-domain reference	10	0	0
Mean ten (secondary)	Legacy point	20	10	0
	Balanced point	20	10	0
	Calibrated margin	11	2	1
	Full-domain reference	10	0	0

S10.1 A FIXED PARTIAL-BALANCE CERTIFICATE

The six-row rule tested below is a sufficient certificate for a shared-domain direction, not a predictor of external model quality. Its row set was selected on the earlier KTH development outputs and frozen before the new-subject forecasts. Let $Aw = b$ contain normalization and all endpoint-balance equations, and let $z = w - \tau\mathbf{1}$. For each clip, an exact optimal dual y_i gives a nonnegative reduced cost $r_i = c_i - A^\top y_i$ and local minimum ℓ_i. Every fully feasible shared weight then satisfies

$$\bar{c}^\top w = \bar{\ell} + \bar{r}^\top z.$$

Keeping normalization and only a fixed subset S of balance rows defines a relaxation

$$\mathcal{Z}_S = \{z \geq 0 :\ A_S z = b_S - \tau A_S \mathbf{1}\}.$$

Consequently, $\bar{\ell} + \min_{z \in \mathcal{Z}_S} \bar{r}^\top z$ is a lower bound on the full shared minimum. Applying the same construction to $-c_i$ gives an upper bound. A direction is reported only when this outer interval excludes zero. These are deterministic guarantees conditional on the finite domain; empirical quantities are applicability, abstention and missed robust directions.

For construction B, S contains the six endpoint rows E012, E021, E120, E210, L012 and L021, plus normalization. This is the first lexicographic minimum-cardinality set identified for the earlier ConvLSTM B mean-ten example, conditional on its frozen local duals. It is not claimed to be an invariant or universally minimum set. For A, all six endpoint rows are retained, so the certificate recovers the full domain and does not test constraint reduction. Both signs and both predeclared readouts are evaluated. Local LPs remain necessary; neither an optimization-free method nor a speed improvement is claimed.

What transfers. Each of the five subjects has four robust and four ambiguous cells across the two artifacts, constructions and readouts. Balanced and legacy point policies therefore each make four unsupported domain-wide assertions per subject if their signs are generalized beyond their selected weights. The threshold overstates one cell for subject 18 and one for subject 22; it also withholds one robust cell for subject 22. At the primary readout it exactly matches full-domain reporting. The pooled margin rule reports four directions, one unsupported, and misses one robust direction; these are separate descriptive pooled results, not new replication counts.

Table 12: All 20 subject-level cells for the primary first-frame readout. Exact classifications use unrounded endpoints. The full and fixed-row predictions coincide in these cells.

Unit	Artifact/family	Interval (10^{-5})	Margin (%)	Legacy	Balanced	Threshold	Full/rows
18	SimVP/A	[4.458,4.601]	13.661	+	+	+	+
20	SimVP/A	[4.964,5.690]	11.080	+	+	+	+
22	SimVP/A	[3.407,4.274]	11.085	+	+	+	+
24	SimVP/A	[4.084,5.100]	13.570	+	+	+	+
25	SimVP/A	[3.745,4.401]	13.054	+	+	+	+
18	SimVP/B	[-1.951,4.332]	2.119	–	+	?	?
20	SimVP/B	[-3.224,5.345]	1.380	–	+	?	?
22	SimVP/B	[-1.848,2.952]	0.951	–	+	?	?
24	SimVP/B	[-2.167,4.513]	1.825	–	+	?	?
25	SimVP/B	[-2.517,3.273]	0.422	–	+	?	?
18	ConvLSTM/A	[0.448,1.284]	7.354	+	+	+	+
20	ConvLSTM/A	[0.745,1.828]	5.685	+	+	+	+
22	ConvLSTM/A	[0.348,1.111]	5.020	+	+	+	+
24	ConvLSTM/A	[0.571,1.617]	8.295	+	+	+	+
25	ConvLSTM/A	[0.516,1.179]	7.216	+	+	+	+
18	ConvLSTM/B	[-1.595,1.652]	1.093	–	+	?	?
20	ConvLSTM/B	[-3.013,1.835]	1.305	–	–	?	?
22	ConvLSTM/B	[-2.210,1.564]	0.689	–	–	?	?
24	ConvLSTM/B	[-2.079,2.159]	0.570	–	+	?	?
25	ConvLSTM/B	[-1.405,1.367]	0.783	–	+	?	?

Table 13: All 20 subject-level cells for the secondary mean-ten readout. Exact classifications use unrounded endpoints. The full and fixed-row predictions coincide in these cells.

Unit	Artifact/family	Interval (10^{-5})	Margin (%)	Legacy	Balanced	Threshold	Full/rows
18	SimVP/A	[5.053,5.430]	8.923	+	+	+	+
20	SimVP/A	[6.588,6.993]	7.501	+	+	+	+
22	SimVP/A	[5.848,7.101]	10.880	+	+	+	+
24	SimVP/A	[4.027,4.965]	8.103	+	+	+	+
25	SimVP/A	[3.460,3.524]	6.584	+	+	+	+
18	SimVP/B	[-1.739,4.787]	1.585	–	+	?	?
20	SimVP/B	[-7.918,10.175]	0.816	–	+	?	?
22	SimVP/B	[-2.692,4.308]	0.468	–	+	?	?
24	SimVP/B	[-1.938,5.793]	2.458	–	+	?	?
25	SimVP/B	[-2.536,3.790]	1.226	–	+	?	?
18	ConvLSTM/A	[0.447,1.220]	5.996	+	+	+	+
20	ConvLSTM/A	[-0.803,0.506]	0.444	–	–	?	?
22	ConvLSTM/A	[0.212,1.020]	3.681	+	+	?	+
24	ConvLSTM/A	[0.693,1.489]	7.523	+	+	+	+
25	ConvLSTM/A	[0.332,0.875]	5.038	+	+	+	+
18	ConvLSTM/B	[-3.475,1.345]	5.323	–	–	–	?
20	ConvLSTM/B	[-8.492,-1.549]	13.373	–	–	–	–
22	ConvLSTM/B	[-4.170,0.960]	6.726	–	–	–	?
24	ConvLSTM/B	[-3.789,3.404]	2.169	–	–	?	?
25	ConvLSTM/B	[-2.200,1.431]	2.091	–	–	?	?

Table 14: All eight pooled new-cohort cells, averaging the 21 eligible clips equally. These reuse the subject forecasts and are not additional independent observations.

Unit	Artifact/family	Interval (10^{-5})	Margin (%)	Legacy	Balanced	Threshold	Full/rows
first	SimVP/A	[4.055,4.732]	12.502	+	+	+	+
mean ten	SimVP/A	[4.903,5.209]	8.264	+	+	+	+
first	SimVP/B	[-2.316,3.921]	1.278	–	+	?	?
mean ten	SimVP/B	[-2.805,5.089]	1.278	–	+	?	?
first	ConvLSTM/A	[0.514,1.362]	6.649	+	+	+	+
mean ten	ConvLSTM/A	[0.238,1.033]	3.784	+	+	?	+
first	ConvLSTM/B	[-1.922,1.677]	0.040	–	+	?	?
mean ten	ConvLSTM/B	[-3.722,0.910]	6.459	–	–	–	?

Computational scope. Inference took 15.5 seconds for SimVP and 74.7 seconds for ConvLSTM in the bounded local GPU workers; these are run receipts, not comparative architecture benchmarks. On the analysis CPU run, the 48 subject/pooled exact group optimizations took 0.342 seconds in total, whereas the partial-row computations took 1.287 seconds, excluding the local certificates. Pair readouts plus local certificates took a further 4.655 seconds. The scalar legacy and balanced group summaries took 0.065 and 0.060 seconds respectively; the threshold adds only a scalar comparison after the balanced summary. These measurements do not support a speed claim for partial rows. Their role here is testing whether a fixed subset retains robust directions. The B subset is informative on only one robust new-subject cell; the A rule keeps every balance row by definition.

S10.2 Post-hoc stratified threshold baselines

After inspecting the prospective results, we evaluated three natural calibration strata to test whether the single global threshold was unnecessarily restrictive. This analysis is post hoc and does not replace the prespecified global comparison. Every threshold is still fitted only to the four old subjects, by taking the largest unsupported development margin within its stratum, or zero if none exists, and applying a strict greater-than rule. No new-subject margin selects a threshold. All three specified levels are reported; their development groups contain 16, 8 and 4 correlated cells respectively, always from only four subjects.

Table 15: All post-hoc stratified baselines on the same 40 subject cells. U is unsupported reports; Miss is withheld robust directions. Every variant ties full-domain reporting on the 20 primary first-frame cells (10 reports, zero U, zero Miss). Differences below are confined to the secondary mean-ten readout.

Old-data calibration strata	Reports	U	Miss
One global threshold (prespecified)	21	2	1
Construction	22	2	0
Artifact $\times$ construction	22	2	0
Artifact $\times$ construction $\times$ readout	24	4	0
Full interval	20	0	0

The family-aware baselines recover the robust A direction withheld by the global threshold, so that miss is not evidence that all scalar calibration is inadequate. They retain the two unsupported B reports. Finer stratification increases overstatement in this finite test. This is evidence about the listed old-data calibration rules, not an impossibility result for every scalar, learned or externally calibrated reporting policy.

S11 Weight-Floor and Structural-Readout Sensitivity

This is a post-hoc saved-output analysis designed after the five-subject study. It adds no training, predictor calls, independent subjects or target-frame access. All 21 originally admitted clips are retained, with the original 30-clip denominator. Two models, two constructions and two horizons yield 40 subject cells and eight separate pooled cells per readout. Eleven equally spaced floor fractions from zero to one are evaluated for every cell; none is selected as a better default after observing the outcomes. There are 1,056 grid intervals and a complete set of rational endpoint certificates.

Nested domains and transition brackets. For $0 \le \alpha_1 \le \alpha_2 \le 1$, $\mathcal{P}_{\alpha_2 t_*} \subseteq \mathcal{P}_{\alpha_1 t_*}$, so the minimum is nondecreasing and the maximum nonincreasing. Strict one-sidedness cannot reverse sign along this path. If the zero-floor interval is one-sided, the direction holds throughout; if the maximum-floor interval is not strictly one-sided, none of the smaller floors can establish a strict direction. Otherwise, ten rational bisections maintain a non-one-sided left endpoint and a strictly one-sided right endpoint. The resulting width is $1/1024$ in α; it is a bracket, not an exact critical value. All intermediate endpoints are certified. The $\alpha = 0.5$ pixel-L_1 intervals reproduce all 48 earlier subject/pooled intervals (96 endpoints) exactly. The zero-floor domain is the nonnegative closure and can omit edges; it is not the same positive-support requirement as any positive floor.

Structural readout. We compute $1 - \mathrm{SSIM}$ per grayscale future frame, then take the first frame or average ten scores before constructing edge costs. We use scikit-image 0.26.0, float64, nominal data range 1, Gaussian weights with $\sigma = 1.5$, population covariance, $K_1 = 0.01$ and $K_2 = 0.03$, following the documented Wang-compatible configuration. No clipping or per-image rescaling is applied. The archived prediction ranges are [0.06839,1.02929] for SimVP and [0.00809,1.00671] for ConvLSTM; respectively 10,749 and 164 of 68,648,960 values exceed the nominal unit interval. This explicit convention preserves the same saved outputs as the pixel comparison. It is a structural readout, not a human-quality ground truth. Eligibility and exact-distance controls remain defined in input pixel L_1, not input SSIM. The $Q/L/D$ inequalities are not asserted for SSIM; the LP remains valid because fixed pair scores enter linearly in the weights.

Complete floor grid. Table 16 gives positive/negative/ambiguous counts. Each row summarizes ten cells sharing five subjects, not ten independent samples. Subject and pooled interval values, costs, feasible endpoint weights and duals are retained in the companion RANGES.json. L1 critical brackets are retained in CRITICAL_BRACKETS.json; structural pair scores are in SSIM_PAIRS.json. The executable entry points prepare_costs.py and analyze.py recompute readouts and ranges from saved forecasts without loading a predictor. The independent verification entry point checks the stored arithmetic without rerunning an optimizer.

Table 16: Complete subject-level counts P/N/A (positive/negative/ambiguous), ten cells per row. Columns are floor fractions α.

Readout / family / horizon	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
L1/A/first	9/0/1	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0
L1/B/first	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	2/0/8	5/2/3
L1/A/mean	8/0/2	8/0/2	8/0/2	9/0/1	9/0/1	9/0/1	9/0/1	9/0/1	9/0/1	9/1/0	9/1/0
L1/B/mean	0/0/10	0/0/10	0/0/10	0/1/9	0/1/9	0/1/9	0/1/9	0/2/8	0/3/7	3/3/4	5/5/0
S/A/first	5/0/5	5/0/5	5/0/5	8/0/2	8/0/2	9/0/1	9/0/1	10/0/0	10/0/0	10/0/0	10/0/0
S/B/first	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/0/10	0/1/9	0/2/8	1/8/1
S/A/mean	7/2/1	7/2/1	7/2/1	7/2/1	7/2/1	7/2/1	7/2/1	7/2/1	8/2/0	8/2/0	8/2/0
S/B/mean	0/1/9	0/1/9	0/1/9	0/1/9	0/1/9	0/1/9	0/3/7	0/4/6	0/4/6	2/4/4	3/5/2

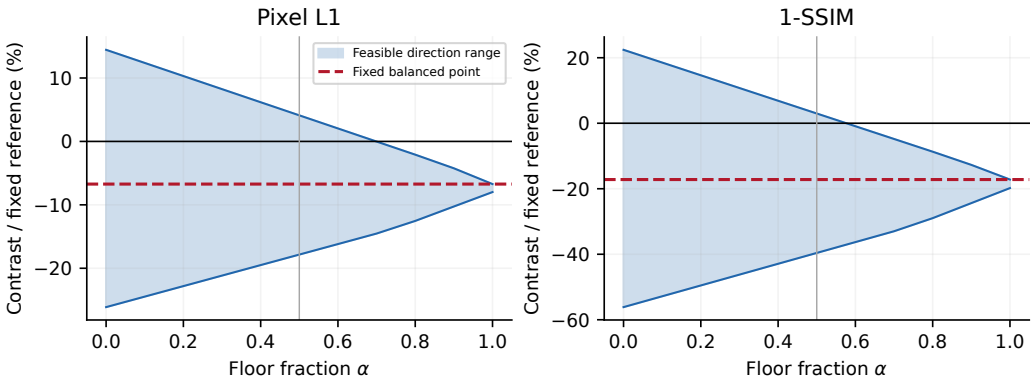

Figure 7: Fixed running example: subject 22, ConvLSTM B, mean ten. Each contrast is divided by that readout's fixed positive balanced-reference mean; normalization changes display units, not the direction. The point weighting is fixed and feasible throughout, while larger floors restrict the admissible alternatives. Lines connect evaluated grid points; only the separately certified L1 bisection brackets localize a transition. Both half-floor intervals cross zero despite negative point estimates.

Complete category changes at the half-floor. Among the 40 subject cells, 36 categories agree between pixel L1 and structural disagreement. All four differences occur for ConvLSTM A: subject 22 first and subject 18 mean ten change positive to ambiguous; subject 20 mean ten changes ambiguous to negative; subject 25 mean ten changes positive to negative. All 20 construction-B categories agree. In the eight pooled comparisons, seven agree and ConvLSTM A mean ten changes positive to negative. Pooled rows reuse the same forecasts and are not additional independent replications.

Table 17: All 48 pixel-L1 floor-path classifications. T: one-sided throughout; A: not strictly one-sided even at maximum floor; other entries are the certified transition bracket, with endpoints divided by 1024. The direction after transition is shown by + or –.

Artifact	Family	Horizon	18	20	22	24	25	Pooled
SimVP	A	first	T+	T+	T+	T+	T+	T+
SimVP	A	mean10	T+	T+	T+	T+	T+	T+
SimVP	B	first	884–885+	935–936+	948–949+	905–906+	997–998+	938–939+
SimVP	B	mean10	855–856+	979–980+	972–973+	832–833+	915–916+	904–905+
ConvLSTM	A	first	T+	T+	44–45+	T+	T+	T+
ConvLSTM	A	mean10	T+	908–909−	242–243+	T+	T+	205–206+
ConvLSTM	B	first	A	932–933−	981–982−	A	A	A
ConvLSTM	B	mean10	808–809−	292–293−	715–716−	963–964−	922–923−	707–708−

The structural readout uses 2,520 unique forecast pairs and 25,200 frame comparisons, taking 29.6 seconds in this CPU run. Grid optimization and critical brackets took 12.6 seconds. These are execution receipts on small fixed graphs, not scaling benchmarks. No learned perceptual model or GPU was used.

S11.1 Earlier readout details

ConvLSTM A is positive at the first frame and negative at mean ten throughout either weight domain. Complete post-hoc per-step product ranges are positive at steps 1–2, cross zero at steps 3–4, and are negative at steps 5–10. SimVP A is negative only at step 3. These patterns locate the horizon dependence without identifying a causal recurrent mechanism. Deleting some subjects also removes ConvLSTM A's small positive first-frame product bound. Complete step and omission grids remain in Appendix S26.

Aggregation can hide response even without a sign comparison. For the initial unmasked first-future fixed pair P_{102}, P_{201}, aggregate-error change retains 54.2% of prediction disagreement for SimVP and 18.8% for ConvLSTM, measured as ratios of mean D to mean Q. The ordering of input histories, the reference weighting and the readout are therefore separate parts of the evaluator's specification; none is supplied by a scalar forecast error alone.

When computation is limited. Reuse saved predictions and solve each declared domain once per readout. If only point contrasts are affordable, report them as point-specific responses and withhold claims over all admissible references. The calibrated margin rule tied the full-domain decisions on the primary readout here, but made two unsupported reports and missed one robust direction at mean ten (Table 11); it is not a validated substitute across horizons. A reduced-row certificate can establish a direction when it succeeds, but the present B evidence covers only one robust cell and no runtime speedup is established. Prioritize full-domain checks for claims that require reference-invariant direction.

Earlier KTH point scales. For scale, the ConvLSTM first-frame default differences are $+3.8\%$ (A) and -8.5% (B) of their respective reference means. The original input-defined B witness reverses the default ordering on SimVP but not ConvLSTM (0.9151 to 0.9824); failure of that one witness to reverse is compatible with a two-sided complete range.

S12 Subject-Composition Sensitivity

This post-hoc analysis separates reference-choice ambiguity from sensitivity to the composition of the observed subject set. Each of the five subjects contributes two artifact cells per grid setting and eight correlated artifact/construction/horizon cells for aggregate path and agreement summaries. We enumerate all $5^5 = 3125$ ordered resamples of five subjects with replacement, carrying each subject's full vector together. All readouts and floors use paired resamples. The estimand is the equal-subject mean fraction of cells in a category, not a reoptimized pooled shared-weight contrast. Eligibility, costs and certified individual-subject intervals stay fixed; no predictor or optimizer is run. Clip-level

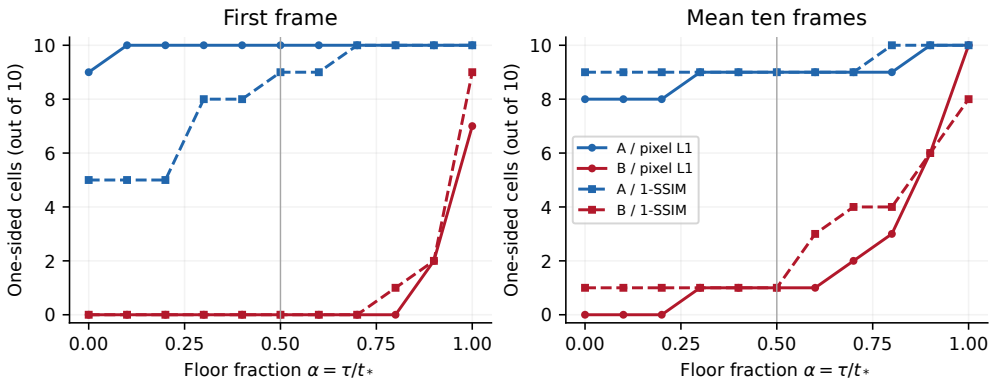

Figure 8: Complete floor grid on the five new study subjects, with ten correlated cells per construction/readout curve within each horizon panel. Curves connect evaluated grid points; they are not confidence bands. Increasing the floor removes admissible reference weights. The vertical line marks the previously declared half-floor, not an optimized choice.

composition is not resampled: each subject classification uses its full eligible clip set; resampling clips would require new optimization.

Reported intervals are the central 95% of the empirical cluster-resampling distribution, using inverse-CDF order statistics 79 and 3047. They are sensitivity summaries, not calibrated population confidence intervals, and no simultaneous coverage is claimed across the 108 summaries: the subjects were not randomly sampled from a specified population, and five clusters cannot reveal unobserved heterogeneity. Degenerate intervals show identical observed contributions, not population certainty. Leave-one-subject-out estimates provide a second, transparent composition check. Full exact fractions and all five omission values accompany RESULTS.json; displayed values below are percentages. An independent multinomial enumeration of 126 count vectors reproduces all 108 summaries.

Table 18: Subject-composition sensitivity. Empirical resampling intervals and leave-one-out extrema describe observed-cluster composition only.

Summary	Estimate	Resampling interval	Leave-one-out range
Ambiguous at maximum floor	7.5	[2.5, 12.5]	[6.2, 9.4]
One-sided throughout	42.5	[32.5, 50.0]	[40.6, 46.9]
Requires positive floor	50.0	[37.5, 65.0]	[43.8, 53.1]
Half-floor readout agreement	90.0	[87.5, 95.0]	[87.5, 90.6]

Observed subject contributions.

Summary	18	20	22	24	25
Ambiguous at maximum floor	12.5	0.0	0.0	12.5	12.5
One-sided throughout	50.0	37.5	25.0	50.0	50.0
Requires positive floor	37.5	62.5	75.0	37.5	37.5
Readout agreement	87.5	87.5	87.5	100.0	87.5

Complete L1 floor grid. Degenerate intervals reflect identical observed subject contributions, not zero population uncertainty.

Family/horizon	α	Estimate	Resampling interval	LOO min	LOO max
A/first	0.0	90.0	[70.0, 100.0]	87.5	100.0
A/first	0.1	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.2	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.3	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.4	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.5	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.6	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.7	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.8	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.9	100.0	[100.0, 100.0]	100.0	100.0
A/first	1.0	100.0	[100.0, 100.0]	100.0	100.0
A/mean10	0.0	80.0	[60.0, 100.0]	75.0	87.5
A/mean10	0.1	80.0	[60.0, 100.0]	75.0	87.5
A/mean10	0.2	80.0	[60.0, 100.0]	75.0	87.5
A/mean10	0.3	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.4	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.5	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.6	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.7	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.8	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.9	100.0	[100.0, 100.0]	100.0	100.0
A/mean10	1.0	100.0	[100.0, 100.0]	100.0	100.0
B/first	0.0	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.1	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.2	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.3	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.4	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.5	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.6	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.7	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.8	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.9	20.0	[0.0, 40.0]	12.5	25.0
B/first	1.0	70.0	[50.0, 90.0]	62.5	75.0
B/mean10	0.0	0.0	[0.0, 0.0]	0.0	0.0
B/mean10	0.1	0.0	[0.0, 0.0]	0.0	0.0
B/mean10	0.2	0.0	[0.0, 0.0]	0.0	0.0
B/mean10	0.3	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.4	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.5	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.6	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.7	20.0	[0.0, 40.0]	12.5	25.0
B/mean10	0.8	30.0	[10.0, 50.0]	25.0	37.5
B/mean10	0.9	60.0	[50.0, 80.0]	50.0	62.5
B/mean10	1.0	100.0	[100.0, 100.0]	100.0	100.0

Complete SSIM floor grid. Degenerate intervals reflect identical observed subject contributions, not zero population uncertainty.

Family/horizon	α	Estimate	Resampling interval	LOO min	LOO max
A/first	0.0	50.0	[50.0, 50.0]	50.0	50.0
A/first	0.1	50.0	[50.0, 50.0]	50.0	50.0
A/first	0.2	50.0	[50.0, 50.0]	50.0	50.0
A/first	0.3	80.0	[60.0, 100.0]	75.0	87.5
A/first	0.4	80.0	[60.0, 100.0]	75.0	87.5
A/first	0.5	90.0	[70.0, 100.0]	87.5	100.0
A/first	0.6	90.0	[70.0, 100.0]	87.5	100.0
A/first	0.7	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.8	100.0	[100.0, 100.0]	100.0	100.0
A/first	0.9	100.0	[100.0, 100.0]	100.0	100.0
A/first	1.0	100.0	[100.0, 100.0]	100.0	100.0
A/mean10	0.0	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.1	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.2	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.3	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.4	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.5	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.6	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.7	90.0	[70.0, 100.0]	87.5	100.0
A/mean10	0.8	100.0	[100.0, 100.0]	100.0	100.0
A/mean10	0.9	100.0	[100.0, 100.0]	100.0	100.0
A/mean10	1.0	100.0	[100.0, 100.0]	100.0	100.0
B/first	0.0	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.1	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.2	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.3	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.4	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.5	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.6	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.7	0.0	[0.0, 0.0]	0.0	0.0
B/first	0.8	10.0	[0.0, 30.0]	0.0	12.5
B/first	0.9	20.0	[0.0, 40.0]	12.5	25.0
B/first	1.0	90.0	[70.0, 100.0]	87.5	100.0
B/mean10	0.0	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.1	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.2	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.3	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.4	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.5	10.0	[0.0, 30.0]	0.0	12.5
B/mean10	0.6	30.0	[10.0, 50.0]	25.0	37.5
B/mean10	0.7	40.0	[20.0, 50.0]	37.5	50.0
B/mean10	0.8	40.0	[20.0, 50.0]	37.5	50.0
B/mean10	0.9	60.0	[50.0, 80.0]	50.0	62.5
B/mean10	1.0	80.0	[60.0, 100.0]	75.0	87.5

Paired floor changes. Changes are percentage points in one-sided prevalence; they describe stricter reference domains, not improved forecasting.

Readout	Family/horizon	α change	Estimate	Resampling interval
L1	A/first	0.0 to 0.5	10.0	[0.0, 30.0]
L1	A/first	0.5 to 1.0	0.0	[0.0, 0.0]
L1	A/mean10	0.0 to 0.5	10.0	[0.0, 30.0]

Readout	Family/horizon	α change	Estimate	Resampling interval
L1	A/mean10	0.5 to 1.0	10.0	[0.0, 30.0]
L1	B/first	0.0 to 0.5	0.0	[0.0, 0.0]
L1	B/first	0.5 to 1.0	70.0	[50.0, 90.0]
L1	B/mean10	0.0 to 0.5	10.0	[0.0, 30.0]
L1	B/mean10	0.5 to 1.0	90.0	[70.0, 100.0]
SSIM	A/first	0.0 to 0.5	40.0	[20.0, 50.0]
SSIM	A/first	0.5 to 1.0	10.0	[0.0, 30.0]
SSIM	A/mean10	0.0 to 0.5	0.0	[0.0, 0.0]
SSIM	A/mean10	0.5 to 1.0	10.0	[0.0, 30.0]
SSIM	B/first	0.0 to 0.5	0.0	[0.0, 0.0]
SSIM	B/first	0.5 to 1.0	90.0	[70.0, 100.0]
SSIM	B/mean10	0.0 to 0.5	0.0	[0.0, 0.0]
SSIM	B/mean10	0.5 to 1.0	70.0	[50.0, 90.0]

S13 Complementary Content-Preserving Reversal Diagnostic

We reuse the thirty already observed KTH histories (five subjects, six actions) and both released artifacts. Alongside native history $[0, 1, \ldots, 9]$, we evaluate full reversal $[9, 8, \ldots, 0]$ and anchored interior reversal $[0, 4, 3, 2, 1, 5, 6, 7, 8, 9]$. Both preserve the frame multiset without repetition or omission. The interior arm also preserves the first frame and last five frames. Actual input checks find that neither arm preserves the native directed-transition multiset on any of the thirty clips. Per-clip frame and directed-transition checks are retained in REVERSAL_FEASIBILITY.json and independently recounted in REVERSAL_RESULT_REVIEW.json. No original A/B graph-eligibility exclusion is applied to this different diagnostic.

The sixty native forecasts are reused after cache verification. One native engineering forward per artifact matches its saved prediction bitwise; the two new arms require 120 scientific forwards without training. The arm definitions, first-frame primary Q, mean-ten secondary Q, and equal-subject aggregation were specified before the new outputs. This is a new diagnostic on an already analyzed cohort, not an independent dataset or checkpoint replication. No future targets are read: these responses cannot establish forecast accuracy or semantic temporal ability.

Table 22 reports native-to-transformed prediction disagreement and full-minus-interior paired differences. Each subject contributes six clip means; all five subject rows are retained below. Values are raw pixel-L_1 disagreement multiplied by 10^3. The two arms differ in perturbation magnitude and several temporal properties, so their difference does not isolate recency or directionality. This three-history set alone supplies no matched/reference family for testing LP ambiguity.

Table 22: Content-preserving reordering responses ($10^3 Q$). These are changes in predictions, not prediction errors.

Artifact	Horizon	Full reversal	Interior reversal	Difference
SimVP	first	18.329	0.793	17.536
SimVP	mean10	16.192	1.305	14.887
ConvLSTM	first	20.565	0.483	20.082
ConvLSTM	mean10	17.671	0.817	16.854

Mean input perturbation over ten frames is 0.017809 for full reversal and 0.003515 for interior reversal. These magnitudes are reported rather than assumed matched.

Artifact	Horizon	Subject	Full	Interior	Difference
SimVP	first	18	10.374	0.541	9.833
SimVP	first	20	23.109	0.779	22.330
SimVP	first	22	15.561	0.821	14.739
SimVP	first	24	23.757	0.958	22.800

Artifact	Horizon	Subject	Full	Interior	Difference
SimVP	first	25	18.841	0.865	17.977
SimVP	mean10	18	10.223	0.978	9.245
SimVP	mean10	20	21.949	1.306	20.642
SimVP	mean10	22	12.840	1.319	11.521
SimVP	mean10	24	21.100	1.500	19.600
SimVP	mean10	25	14.847	1.421	13.426
ConvLSTM	first	18	12.626	0.270	12.357
ConvLSTM	first	20	23.421	0.624	22.796
ConvLSTM	first	22	20.284	0.525	19.760
ConvLSTM	first	24	22.313	0.499	21.814
ConvLSTM	first	25	24.180	0.499	23.681
ConvLSTM	mean10	18	10.225	0.418	9.808
ConvLSTM	mean10	20	20.808	1.225	19.583
ConvLSTM	mean10	22	18.487	0.872	17.615
ConvLSTM	mean10	24	20.051	0.861	19.190
ConvLSTM	mean10	25	18.785	0.710	18.075

S13.1 THIRD-PREDICTOR REPLICATION

We selected the official OpenSTL KTH20 PredRNN checkpoint before running the fixed R26 inputs. Its four 128-channel recurrent layers propagate a shared spatiotemporal memory between layers. The released checkpoint has 23,568,384 parameters. We use its official patch-size-two interface, ten history frames and twenty autoregressive predictions, retaining the first ten. Future placeholders are zero and the scheduled-sampling mask is zero; no target enters inference. Two synthetic checks precede 419 unique scientific predictions in FP32, batch size one. The same 21 eligible clips, five subjects, graphs and half-floor domains are retained.

The table gives every half-floor shared interval in raw matched-minus-reference pixel disagreement. Pooled rows average the 21 eligible clips equally. Exact endpoints, zero/half/maximum-floor shared intervals and half-floor product intervals are stored in the accompanying results. All endpoints have rational primal/dual certificates. The standalone replay package verifies 72 shared comparisons across three predictors (60 subject cells and 12 pooled cells), comprising 144 endpoint certificates; its optional solver mode recomputes them from saved costs.

Table 24: Complete PredRNN half-floor shared ranges.

Family	Horizon	Subject	Lower	Upper	Direction
A	first	18	-5.04624e-05	-3.44865e-05	negative
A	first	20	-5.31892e-05	-4.11319e-05	negative
A	first	22	-6.62266e-05	-4.37183e-05	negative
A	first	24	-5.37968e-05	-3.85931e-05	negative
A	first	25	-1.42102e-05	-8.90655e-06	negative
A	first	pooled	-4.4132e-05	-3.06679e-05	negative
A	mean10	18	-0.000260908	-0.000229161	negative
A	mean10	20	-0.000113328	-9.81372e-05	negative
A	mean10	22	-0.000173036	-0.000139156	negative
A	mean10	24	-9.29989e-05	-7.16626e-05	negative
A	mean10	25	-4.21285e-05	-3.56972e-05	negative
A	mean10	pooled	-0.000126426	-0.000110195	negative
B	first	18	-0.000111017	-2.84018e-05	negative
B	first	20	-0.000150284	-3.26968e-05	negative
B	first	22	-0.00015616	-3.68064e-05	negative
B	first	24	-9.62075e-05	-2.65146e-05	negative
B	first	25	-6.14987e-05	-7.23754e-07	negative
B	first	pooled	-0.000106948	-2.48928e-05	negative
B	mean10	18	-0.000353626	-0.000134248	negative

Family	Horizon	Subject	Lower	Upper	Direction
B	mean10	20	-0.000326181	-0.000130107	negative
B	mean10	22	-0.00034717	-0.000136742	negative
B	mean10	24	-0.000221204	-8.13944e-05	negative
B	mean10	25	-0.000141794	-3.02663e-05	negative
B	mean10	pooled	-0.00025734	-9.88149e-05	negative

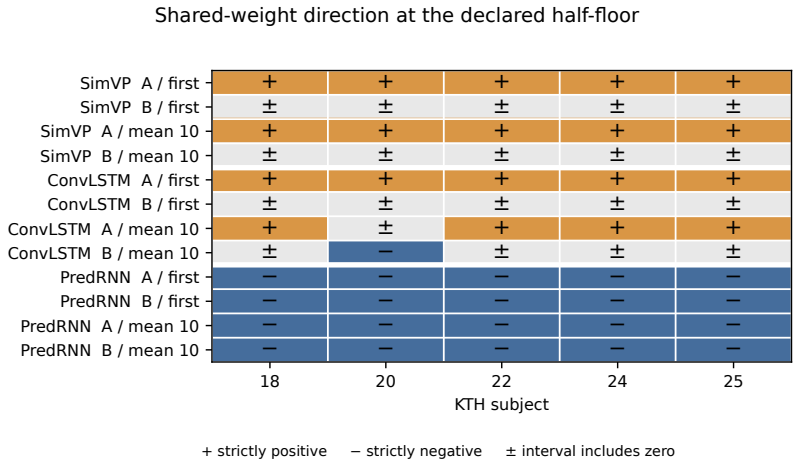

Figure 9: All 60 subject cells at the fixed half-floor, using the same 21 eligible clips out of 30. Signs describe matched-minus-reference pixel disagreement over every shared weight: $+$ or $-$ is strictly one-sided; $\pm$ includes zero. Columns are five subject clusters, with correlated cells within each column. PredRNN changes the model pattern; its zero-floor B ambiguities are reported in the text.

S13.2 Cross-predictor ordering under common reference rules

This saved-output analysis uses all three predictors on the same 21 eligible KTH clips and five subjects. For each pair (f, g), construction and horizon, its estimand is $G_{fg}(w) = C_f(w) - C_g(w) = (\bar{c}_f - \bar{c}_g)^\top w$. We optimize this difference directly on the shared domain, rather than subtract independently optimized model intervals. The three floors are $0, t_*/2, t_*$. Equal clip averaging within each subject and over the pooled cohort gives 216 ranges, with 432 exact primal/dual endpoint certificates.

For magnitude reporting, fix the inherited input-only witness w_0 and define $S_{fg} = [M_f(w_0) + R_f(w_0) + M_g(w_0) + R_g(w_0)]/4 > 0$. Report the raw interval and its endpoints divided by this constant scale. A crossing interval has two-sided magnitude $\min(-\ell, h)/S_{fg}$. The pre-analysis descriptive thresholds are 1%, 5%, and 10%; none is a validated utility threshold. The independent model-normalization diagnostic evaluates $C_f(w)/[(M_f(w) + R_f(w))/2]$ at the two raw endpoint witnesses. It tests whether those two witnesses also reverse the relative-score order; it does not optimize or bound that fractional score globally.

At zero floor, 13/60 subject and 2/12 pooled comparisons cross zero. At the half-floor, these counts are 9/60 and 1/12; at the maximum floor, both are zero. All half-floor crossings compare SimVP with ConvLSTM on B: first-frame for all five subjects, mean-ten for subjects 20, 22, 24, 25, and pooled first-frame. The five first-frame subject intervals have two-sided magnitude above 5%; none of the nine exceeds 10%. Six of the nine subject crossings and the pooled crossing also reverse the relative score at the same witnesses. At zero floor those counts are 10/13 and 1/2. Each reported comparison is correlated with other cells from the same subject.

The illustration rule was fixed before this analysis: prefer pooled half-floor first-frame crossings, then other pooled half-floor crossings, then subject first-frame and remaining subject crossings; within a group choose the largest two-sided normalized margin. It selects the single pooled half-floor

crossing, SimVP–ConvLSTM first-frame B. Full subject and pooled costs, witnesses, model values, and certificates accompany the standalone standard-library replay. The below pooled table and figure retain stable as well as reversing comparisons.

Table 25: All pooled cross-predictor ranges. Relative endpoints are percentages of S_{fg}.

Models	Setting	α	Raw gap	Relative gap (%)
SimVP–ConvLSTM	A/first	0	[1.93e-05,4.98e-05]	[7.44,19.19]
SimVP–ConvLSTM	A/first	1/2	[2.69e-05,4.22e-05]	[10.37,16.25]
SimVP–ConvLSTM	A/first	1	[3.46e-05,3.46e-05]	[13.31,13.31]
SimVP–ConvLSTM	A/mean10	0	[3.32e-05,5.52e-05]	[8.22,13.66]
SimVP–ConvLSTM	A/mean10	1/2	[3.87e-05,4.97e-05]	[9.58,12.30]
SimVP–ConvLSTM	A/mean10	1	[4.42e-05,4.42e-05]	[10.94,10.94]
SimVP–ConvLSTM	B/first	0	[-4.59e-05,6.71e-05]	[-15.52,22.71]
SimVP–ConvLSTM	B/first	1/2	[-2.21e-05,3.94e-05]	[-7.47,13.33]
SimVP–ConvLSTM	B/first	1	[4.94e-06,1.17e-05]	[1.67,3.96]
SimVP–ConvLSTM	B/mean10	0	[-2.07e-05,6.44e-05]	[-4.51,14.05]
SimVP–ConvLSTM	B/mean10	1/2	[7.82e-07,4.6e-05]	[0.17,10.02]
SimVP–ConvLSTM	B/mean10	1	[2.38e-05,2.66e-05]	[5.18,5.80]
SimVP–PredRNN	A/first	0	[6.11e-05,0.000102]	[17.63,29.30]
SimVP–PredRNN	A/first	1/2	[7.12e-05,9.15e-05]	[20.55,26.39]
SimVP–PredRNN	A/first	1	[8.13e-05,8.13e-05]	[23.47,23.47]
SimVP–PredRNN	A/mean10	0	[0.00015,0.000188]	[23.34,29.36]
SimVP–PredRNN	A/mean10	1/2	[0.000159,0.000179]	[24.85,27.86]
SimVP–PredRNN	A/mean10	1	[0.000169,0.000169]	[26.35,26.35]
SimVP–PredRNN	B/first	0	[1.35e-06,0.000146]	[0.34,36.51]
SimVP–PredRNN	B/first	1/2	[2.99e-05,0.00011]	[7.45,27.45]
SimVP–PredRNN	B/first	1	[6.35e-05,7.21e-05]	[15.86,18.00]
SimVP–PredRNN	B/mean10	0	[2.52e-05,0.000312]	[3.51,43.54]
SimVP–PredRNN	B/mean10	1/2	[0.000108,0.000258]	[15.11,36.01]
SimVP–PredRNN	B/mean10	1	[0.000193,0.000201]	[26.95,27.97]
ConvLSTM–PredRNN	A/first	0	[4.18e-05,5.18e-05]	[17.95,22.24]
ConvLSTM–PredRNN	A/first	1/2	[4.43e-05,4.93e-05]	[19.02,21.17]
ConvLSTM–PredRNN	A/first	1	[4.68e-05,4.68e-05]	[20.09,20.09]
ConvLSTM–PredRNN	A/mean10	0	[0.000116,0.000133]	[28.54,32.60]
ConvLSTM–PredRNN	A/mean10	1/2	[0.000121,0.000129]	[29.56,31.59]
ConvLSTM–PredRNN	A/mean10	1	[0.000125,0.000125]	[30.57,30.57]
ConvLSTM–PredRNN	B/first	0	[3.18e-06,0.000122]	[1.05,40.54]
ConvLSTM–PredRNN	B/first	1/2	[2.99e-05,9.28e-05]	[9.93,30.81]
ConvLSTM–PredRNN	B/first	1	[5.86e-05,6.04e-05]	[19.44,20.05]
ConvLSTM–PredRNN	B/mean10	0	[1.4e-05,0.000274]	[2.90,56.67]
ConvLSTM–PredRNN	B/mean10	1/2	[9.02e-05,0.000226]	[18.67,46.83]
ConvLSTM–PredRNN	B/mean10	1	[0.00017,0.000174]	[35.13,36.06]

One witness pair across all subjects. An additional post-hoc decomposition applies the selected pooled witnesses unchanged to each subject, with no new optimization. Raw SimVP-minus-ConvLSTM gaps at w_- are $[-2.115, -2.536, -1.535, -2.575, -2.307] \times 10^{-5}$ for subjects 18,20,22,24,25; at w_+ they are $[4.148, 5.219, 3.475, 5.024, 2.748] \times 10^{-5}$. The corresponding relative-score gaps are $[-6.69, -3.93, -3.38, -8.20, -6.46]$ and $[14.90, 12.88, 12.30, 18.58, 10.45]$ percentage points. Hence one common rule pair reverses all five subject orderings, including the model-specific normalization. These weights were obtained from pooled outputs; this decomposition checks recurrence within the observed cohort.

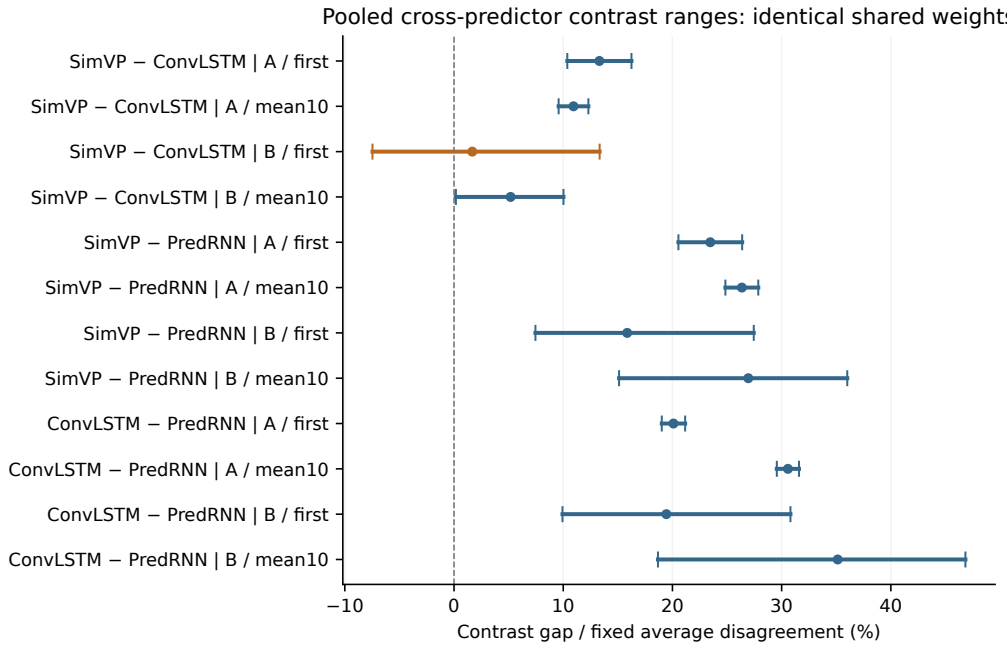

Figure 10: All twelve pooled half-floor intervals. Dots mark the inherited input-defined witness; orange indicates the one cross-zero interval. The denominator is fixed during each optimization.

S13.3 Graph-only rules and fixed-weight sequence confirmation

For each inherited graph, the three objectives are $\max -\sum_e w_e \log w_e$, $\min \sum_e w_e^2$ on the half-floor domain, and $\min \sum_e w_e^2$ on the maximum-floor face. The last objective resolves ties after maximizing the minimum edge mass. The three A weights coincide; all are retained. Numeric optimization is followed by rational feasibility reconstruction. The maximum conditional-gradient gap is below 10^{-8}; logarithmic entropy gradients are numerical. Saved weights satisfy the floor and balance equations exactly. These graph-only objectives were fixed before computing their scores from previously studied outputs.

All 216 model scores and 216 pairwise scores are released in the corrected JSON results and accompanying complete comparison CSV. The single raw disagreement has maximum-entropy, minimum-L2 and maximum-floor gaps 0.821, -1.260 and 2.707 in units of 10^{-6}; the corresponding normalized gaps are 0.569, 0.150 and 0.971 percentage points, all positive. No rule was replaced after observing these results.

The first-frame pooled B endpoints selected in the preceding five-subject analysis remain unchanged for both transfer checks. The earlier-cohort check is retrospective: its forecasts had already been analyzed. The new-sequence check instead fixes its candidate rule, weights, readout and admission test before decoding the new histories and forecasting. It takes exactly the first ten frames of the second annotated subsequence of each of the thirty earlier d1 videos, six actions per subject (18, 20, 22, 24, 25). These histories do not overlap the original first ten frames. All thirty share the required actual-pixel matched pairs, references and exact spatial-distance graph. Each subject contributes six admitted clips. All 300 constructed inputs have new byte content relative to the two earlier KTH input pools.

The released SimVP and ConvLSTM artifacts each make 300 scientific forwards and one engineering forward that reproduces a cached native prediction bitwise. The scientific forecast total becomes 5,717; the two engineering checks are counted separately. No training or target-error scoring is added. Predictions retain ten future frames. For each horizon, clip-average matched and reference responses precede the normalized score $2(M-R)/(M+R)$. Both models use the identical fixed witness in each column. Directional replication means $G(w_-) < 0 < G(w_+)$; a reversal of the opposite orientation

would be recorded separately. First-frame pooled replication is primary, subject decompositions and mean ten are secondary. New segments of previously used videos test temporal-segment transfer, with the existing upstream checkpoint-selection scope unchanged.

Table 26: Every fixed-witness transfer gap. Raw units are 10^{-5}; normalized units are percentage points.

Study/subject	Horizon	Raw w_-	Raw w_+	Norm. w_-	Norm. w_+
Earlier/17	first	-3.953	8.919	-6.29	17.46
Earlier/19	first	-1.938	3.023	-3.93	13.95
Earlier/21	first	-3.919	4.085	-6.00	10.28
Earlier/23	first	-3.233	10.383	-2.81	17.66
Earlier/pooled	first	-3.228	6.801	-4.56	16.06
Earlier/17	mean10	-1.508	8.934	-1.01	12.36
Earlier/19	mean10	1.284	4.927	13.58	20.36
Earlier/21	mean10	1.250	9.470	10.37	20.73
Earlier/23	mean10	2.113	29.434	8.39	33.14
Earlier/pooled	mean10	0.532	12.466	7.09	23.70
New/18	first	-1.440	2.850	-3.91	14.36
New/20	first	-1.139	3.771	-5.20	20.38
New/22	first	-1.853	3.784	-8.46	17.52
New/24	first	-1.511	2.893	-5.45	13.56
New/25	first	-2.463	3.738	-9.04	16.06
New/pooled	first	-1.681	3.407	-6.68	16.19
New/18	mean10	0.332	2.256	11.56	9.97
New/20	mean10	1.813	4.870	14.19	19.89
New/22	mean10	1.048	3.099	7.58	10.68
New/24	mean10	-0.349	3.721	2.04	15.61
New/25	mean10	-0.604	2.570	6.49	9.00
New/pooled	mean10	0.448	3.303	8.14	12.93

S13.4 Content-preserving permutations and independent forecast quality

The extension uses the first annotated twenty-frame window from each of thirty KTH d2 videos: subjects 18, 20, 22, 24 and 25, crossed with the six actions. The first ten frames form the history and the following ten the target. The existing d1 histories determine the descriptor width; no d2 forecast or target error enters that choice. All thirty candidates remain in the native-history quality evaluation.

Input-defined construction. We enumerate the 24 permutations of history positions 1–4, keeping position 0 and positions 5–9 fixed (zero-based indexing). Every original index appears exactly once. We require ten pixel-distinct frames. Each history has nine adjacent-frame mean absolute pixel differences; their histogram uses bins $\lfloor 255d \rfloor$. Matched pairs have identical histograms and reference pairs have different histograms. This controls a quantized distribution of unsigned scalar transition magnitudes; it does not preserve directed transitions or natural motion. Joint edges require exactly equal time-summed per-pixel input-distance maps for the matched and reference pairs.

Domain feasibility and transfer. The initial development test found no common matched-pair support across all eligible clips at any of three fixed widths ($1/255$, $2/255$, $4/255$). The refined design therefore transfers an input-defined algorithm on clip-specific domains. For each graph, we retain its maximal attainable support: edges that can receive positive mass under normalization and endpoint balance. An exact primal–dual certificate proves that all omitted edges have zero attainable mass, and a second certificate proves a strictly positive common floor on the retained support. Removing these forced-zero coordinates preserves the zero-floor domain. Applying half the maximal floor on this support defines the extension's positive-floor domain. It does not reproduce the old positive-floor domain or transfer an unchanged numerical weight vector.

The finest development width meeting the fixed criterion of at least ten clips across at least four subjects is $1/255$ (16 clips across all five subjects). Applied unchanged to the thirty d2 candidates,

the input procedure admits 15 clips (2/5/3/2/3 by subject). Seven candidates contain pixel-identical frames, one has no eligible edges, and seven yield an empty numerical support screen; the last category is not an exact infeasibility proof. All thirty remain available for native-history error comparisons. All admitted histories contribute all 24 permutations; global byte-level deduplication, including native inputs for every candidate, yields 375 inputs per predictor. Eligibility, mappings and the weight objectives are fixed before forecasting.

Independent outcome and decisions. The primary hypothesis is that the larger first-frame raw contrast $C = M - R$ selects the predictor with lower native-history ten-frame mean squared error. SimVP versus ConvLSTM is primary; comparisons involving PredRNN are secondary. This score-to-quality link is tested rather than assumed. We evaluate maximum entropy, minimum squared mass, maximum floor with a minimum-squared-mass tie-break, agreement among those three rules, and the full-domain sign. The latter selects only when the complete raw-gap interval excludes zero. Exact ties abstain.

Native-history MSE uses all ten actual future frames, float64 accumulation and unclipped model outputs. We retain coverage, correct and incorrect choices, quality ties, selected MSE and regret against the same-clip two-model oracle. Each selected subset is compared with both fixed predictors on that identical subset. Two separately fixed fallback policies assign abstentions to each of the two predictors; both are reported. Mean-ten probe scores and normalized fixed-rule comparisons are secondary. Raw linear-program bounds are never treated as normalized-score bounds. The five subjects are clusters, and pooled clips are not an additional independent replicate.

Execution and completeness. The initial local execution completes 375 forecasts for each primary predictor, but loses the CUDA device after 285 saved PredRNN predictions and one further consumed input. Its receipts retain 23 PredRNN native predictions and 11 complete permutation grids. The interrupted execution and all 1,035 saved outputs are preserved without resumption or overwrite; its incomplete grids do not enter the reported remote comparison.

The subsequent remote confirmation evaluates all three artifacts on the same frozen 375 inputs in a single environment: RTX 4090, PyTorch 2.7.1 with CUDA 12.6, NumPy 2.2.5, deterministic FP32 and TF32 disabled. All 1,125 remote scientific forecasts complete. Each artifact first passes a separate seeded-kernel and cached-native engineering check. ConvLSTM and PredRNN reproduce the old native cache bitwise; SimVP differs by at most 4.77×10^{-7} per pixel. The reported content-preserving comparison uses the complete remote outputs for all three predictors, without mixing platforms or pooling the interrupted local run as additional data. Both the inputs and weighting rules remain unchanged.

Table 27: Pure-permutation primary comparison: 15 eligible clips from 30 candidates. MSE ($\times 10^{-3}$) is measured on native ten-frame forecasts. Each fixed-model baseline uses exactly the same decided clips as its policy.

Policy	Decided	Correct	MSE	Fixed SimVP	Fixed ConvLSTM
Maximum entropy	15	7	1.5145	1.5377	1.4042
Minimum squared mass	15	8	1.5114	1.5377	1.4042
Maximum floor	15	7	1.5145	1.5377	1.4042
Three-rule unanimity	14	7	1.5870	1.6119	1.4722
Full-domain sign	4	2	1.4620	1.4172	1.4825

Complete pairwise stability summary.

Pair	Horizon	Crossings /15	Raw disagreement	Normalized disagreement
SimVP–ConvLSTM	first	11	1	0
SimVP–ConvLSTM	mean10	11	0	1
SimVP–PredRNN	first	11	0	2
SimVP–PredRNN	mean10	9	0	0
ConvLSTM–PredRNN	first	8	0	0
ConvLSTM–PredRNN	mean10	9	0	1

Native-history baselines.

Predictor	Cohort	Clips	MSE ($\times 10^{-3}$)
SimVP	All candidates	30	1.5072
SimVP	Probe eligible	15	1.5377
ConvLSTM	All candidates	30	1.3768
ConvLSTM	Probe eligible	15	1.4042
PredRNN	All candidates	30	1.3274
PredRNN	Probe eligible	15	1.3539

Both fixed fallback choices.

Probe horizon	Policy	SimVP fallback MSE	ConvLSTM fallback MSE
first	Maximum entropy	1.4956	1.4319
first	Minimum squared mass	1.4941	1.4303
first	Maximum floor	1.4956	1.4319
first	Three-rule unanimity	1.4956	1.4303
first	Full-domain sign	1.5132	1.3740
mean10	Maximum entropy	1.5135	1.4498
mean10	Minimum squared mass	1.5135	1.4498
mean10	Maximum floor	1.5135	1.4498
mean10	Three-rule unanimity	1.5135	1.4498
mean10	Full-domain sign	1.5132	1.3740

Fallback MSE is in units of 10^{-3} on all thirty candidates, including probe-ineligible clips. Every subject, pair and horizon, along with exact costs and interval certificates, is retained in the accompanying machine-readable remote result files. No preferred fallback is selected after outcomes.

Post-hoc random-choice baseline. For a fixed decided subset, choosing each of SimVP and ConvLSTM with probability one half has expected MSE equal to the mean of their two fixed-model MSEs. This exact expectation requires no simulated choices or new forecasts. For maximum entropy, minimum squared mass and maximum floor on the same 15 primary clips, it is 1.4709724×10^{-3}; for unanimity on 14 clips it is 1.5420388×10^{-3}; for the full-domain policy on four clips it is 1.4498158×10^{-3}. Each is lower than its corresponding policy's MSE. This descriptive comparison, added after review, further limits the tested score-to-quality link; it is not a population significance test.

S13.5 Distance from a fixed rule to a ranking boundary

For a frozen joint-edge distribution w_0 and raw model-gap vector c, write $g_0 = c^\top w_0$. When $g_0 \neq 0$ and a strictly opposite sign is attainable in the existing half-floor domain $\mathcal{P}_{t_*/2}$, define

$$r(w_0, c) = \min_{w \in \mathcal{P}_{t_*/2}} \tfrac{1}{2}\|w - w_0\|_1 \quad \text{subject to} \quad \operatorname{sign}(g_0) c^\top w \leq 0.$$

The optimum reaches the zero boundary. Convex mixing with a strictly opposite-sign feasible point shows that r is the infimum distance to strict reversal; a boundary witness does not itself reverse the sign. Centers already tied, domains that only touch zero and strictly one-sided domains are recorded separately. Maximum-floor centers are evaluated in the same half-floor perturbation domain as the other centers.

Total variation measures redistributed mass of the joint-edge distribution, affecting both matched and reference marginals. Its value depends on the declared coupling representation. It is not a percentage of changed frames or a reference-only perturbation. Each content-preserving clip uses its own graph; weights are not transferred across those graphs.

Verification. Auxiliary variables linearize the absolute values. Gap coefficients are scaled for numerical conditioning, while verification retains exact rational values. Equality repair and convex

interpolation supply an exactly feasible zero-gap witness and hence an upper bound. A rational Lagrangian bound corrects every dual residual over finite boxes for both weights and auxiliary variables, providing a lower bound without assuming numerical dual feasibility. The boxes $w_j \in [t_*/2, 1]$ and $u_j \in [0, 1]$ preserve the optimum because weights are probability distributions. Full-domain endpoint primal–dual certificates independently establish opposite-sign attainability. Brackets of width at most 10^{-7} are reported as tightly certified. For a bracket $[L, U]$ and budget β, $U < \beta$ permits strict reversal, $L \geq \beta$ excludes strict reversal and all other cases remain unresolved. Budgets $0.01, 0.05, 0.10, 0.25$ are specified before this analysis.

Retrospective segment comparison. The shared B graph permits reuse of the identical three central rules on the earlier cohort and its later segments. We match the original 21 admitted videos to their 21 later segments by subject, action, video identity and checksum; their history windows do not overlap. The complete thirty later segments are a secondary comparison with a different composition. We compare earlier radii with later central-sign changes for each of five subjects; pooled results are supplementary, and rules and horizons are correlated. These outputs were previously inspected, so this is a retrospective diagnostic test rather than prospective confirmation. Stability to weight changes alone implies no theorem about stability to changes of video segment.

All 378 comparisons complete with independent certificate verification: 267 finite radii have bound widths at most 10^{-7} and 111 are one-sided. There are 270 content-preserving comparisons across clips, model pairs, horizons and rules, plus 108 shared-graph comparisons across cohorts, subjects or pools, horizons and rules. These analyses are correlated. No new forecasts are generated.

Table 28: Primary content-preserving comparison. Budgets refer to total variation of joint-edge mass. Medians condition on the 11 strictly reversible clips; budget counts use all 15 admitted clips. The remaining four are one-sided.

Center	Median radius	$< 1\%$	$< 5\%$	$< 10\%$	$< 25\%$
Maximum entropy	14.40%	1/15	3/15	5/15	11/15
Minimum squared mass	14.39%	1/15	4/15	5/15	11/15
Maximum floor	10.25%	0/15	2/15	5/15	11/15

Table 29: Retrospective paired-segment diagnostic at a fixed 5% budget. A flag predicts a central-sign change; truth is the observed change between the matched cohorts. Each row has five subjects and no excluded ties. TP/FP/TN/FN are confusion counts, not additional observations.

Horizon	Rule	TP	FP	TN	FN	Always-stable correct
First	Maximum entropy	0	3	2	0	5/5
First	Minimum squared mass	1	2	2	0	4/5
First	Maximum floor	0	1	4	0	5/5
Mean ten	Maximum entropy	0	0	5	0	5/5
Mean ten	Minimum squared mass	0	0	5	0	5/5
Mean ten	Maximum floor	0	0	5	0	5/5

The full budget grid and all-thirty later-segment comparison remain in the machine-readable supplement. At mean ten, no paired subject changes its central sign and the 5% flags make no positive prediction. These observations establish neither a useful predictor of segment shifts nor population-level error rates.

S13.6 Task-grounded temporal evaluations

Connection to published diagnostics. Lotter et al. (2017), Figure 4, inspect predictions on a temporally scrambled sequence. Xu et al. (2018), Section 4.6 and Figure 8, inspect predictions when sequentially trained models receive reverse-order inputs. Both order-specific demonstrations are qualitative. We adapt these diagnostic questions to our three frozen predictors and report task error; the architectures and experimental windows differ from the original studies.

Backward task. Each of the thirty existing d2 windows consists of history $H = (x_0, \ldots, x_9)$ and future $T = (x_{10}, \ldots, x_{19})$. Forward prediction uses $H \to T$. Backward prediction uses $(x_{19}, \ldots, x_{10}) \to (x_9, \ldots, x_0)$. All thirty windows and all three artifacts are retained. The design follows inspection of forward results and acquires ninety new backward predictions. Three additional engineering forwards verify exact agreement with prior outputs in the same remote runtime. Each direction has a last-input-frame baseline: repeat x_9 forward and x_{10} backward. The ratio divides the model's aggregate MSE by the same-direction aggregate persistence MSE. Conditioning and targets both change between directions, so the difference includes target difficulty as well as model behavior.

Anchored-permutation task. On the fifteen clips admitted by the content-preserving construction, we reuse every saved prediction from the 24 anchored permutations. Identity supplies the matched native baseline. The stress risk averages the other 23 permutations uniformly while retaining the original future T as target. No reference weights or outcome-selected permutations enter this average. The last history frame remains fixed. This is a restricted perturbation distribution; its error increase measures the cost of those histories for this prediction task.

Readouts and verification. The primary readout averages squared pixel error over ten predicted frames; first-frame error is secondary. Inputs and outputs are float32, with float64 error accumulation and no output clipping. PSNR is $-10 \log_{10}(\mathrm{MSE})$ at peak one, applied after aggregation. Subject means give equal mass to that subject's clips; the pooled mean gives equal mass to all included clips and is not an additional independent subject. Independent framewise recomputation checks all task errors, persistence baselines, ratios, complete subject/model/horizon grids, and rankings against the saved raw outputs.

Table 30: Primary ten-frame task errors. All MSEs are $\times 10^{-3}$. F/B: forward/backward on all 30 clips; ratios divide by same-direction persistence. N/U: native/uniform23 on the same 15 eligible clips. Pooled rows give equal mass to clips.

Subject	Model	F	B	F/pers.	B/pers.	N	U
18	SimVP	1.369990	0.801498	0.667821	0.527545	1.527329	1.539540
18	ConvLSTM	1.323462	0.717904	0.645141	0.472524	1.617996	1.632903
18	PredRNN	1.297389	0.641032	0.632431	0.421927	1.426863	1.692974
20	SimVP	1.267280	0.587442	0.638437	0.415427	1.462707	1.471907
20	ConvLSTM	1.219765	0.580744	0.614499	0.410690	1.412313	1.410754
20	PredRNN	1.273204	0.574880	0.641421	0.406543	1.473638	1.525061
22	SimVP	2.215226	2.015671	0.915363	1.112211	1.613106	1.615195
22	ConvLSTM	1.617938	1.120397	0.668556	0.618215	1.028895	1.028905
22	PredRNN	1.644877	0.892312	0.679687	0.492362	0.980323	0.999305
24	SimVP	1.562708	0.856042	0.761554	0.604152	2.560501	2.562431
24	ConvLSTM	1.584837	0.887984	0.772338	0.626695	2.292195	2.309776
24	PredRNN	1.495020	0.803097	0.728568	0.566786	2.379895	2.255714
25	SimVP	1.120909	1.076350	0.544459	0.617142	0.912439	0.911720
25	ConvLSTM	1.137786	0.997178	0.552657	0.571747	1.031570	1.036255
25	PredRNN	0.926579	0.925723	0.450067	0.530777	0.795188	0.802024
pooled	SimVP	1.507223	1.067401	0.713160	0.674998	1.537722	1.542948
pooled	ConvLSTM	1.376758	0.860841	0.651429	0.544375	1.404223	1.408974
pooled	PredRNN	1.327414	0.767409	0.628081	0.485291	1.353883	1.395111

Table 31: Secondary first-frame task errors. All MSEs are $\times 10^{-3}$. F/B: forward/backward on all 30 clips; ratios divide by same-direction persistence. N/U: native/uniform23 on the same 15 eligible clips. Pooled rows give equal mass to clips.

Subject	Model	F	B	F/pers.	B/pers.	N	U
18	SimVP	0.390745	0.347633	0.822288	0.731564	0.435213	0.437303
18	ConvLSTM	0.261703	0.238724	0.550730	0.502374	0.295661	0.297449
18	PredRNN	0.202519	0.190539	0.426183	0.400974	0.225237	0.237165
20	SimVP	0.328740	0.279579	0.573521	0.487755	0.373726	0.373789
20	ConvLSTM	0.231956	0.198592	0.404671	0.346465	0.269473	0.269919
20	PredRNN	0.187062	0.150756	0.326350	0.263011	0.216471	0.220824
22	SimVP	1.139005	1.228433	1.206395	1.301115	1.081200	1.083352
22	ConvLSTM	0.410921	0.379156	0.435234	0.401589	0.378093	0.378069
22	PredRNN	0.316462	0.314064	0.335186	0.332646	0.316452	0.318504
24	SimVP	0.293501	0.292814	0.642028	0.640525	0.282168	0.284030
24	ConvLSTM	0.228637	0.212891	0.500140	0.465697	0.177172	0.177642
24	PredRNN	0.186916	0.178569	0.408875	0.390616	0.135632	0.147240
25	SimVP	0.364037	0.410641	0.683412	0.770902	0.302142	0.304344
25	ConvLSTM	0.316322	0.341586	0.593835	0.641263	0.279621	0.280414
25	PredRNN	0.271149	0.304842	0.509031	0.572283	0.241162	0.241624
pooled	SimVP	0.503205	0.511820	0.843639	0.858082	0.496894	0.498314
pooled	ConvLSTM	0.289908	0.274190	0.486039	0.459688	0.284411	0.285015
pooled	PredRNN	0.232821	0.227754	0.390332	0.381837	0.231796	0.236887

Table 32: Raw persistence MSE ($\times 10^{-3}$), on all thirty windows. Each subject contributes six windows. The same baseline is used for all three models.

Subject	First forward	First backward	Mean-ten forward	Mean-ten backward
18	0.475192	0.475192	2.051432	1.519297
20	0.573196	0.573196	1.984975	1.414068
22	0.944139	0.944139	2.420050	1.812309
24	0.457146	0.457146	2.051999	1.416930
25	0.532676	0.532676	2.058757	1.744090
pooled	0.596470	0.596470	2.113442	1.581339

Complete ranking pattern. At mean ten, the forward/backward best model changes only on subjects 20 and 22; subject 25 swaps second and third place. Under uniform23 the best model changes on subjects 18 and 24, each represented by two eligible clips. First-frame rankings in all four conditions are PredRNN, ConvLSTM, then SimVP for every subject and the pool. Thus the local task-ranking changes are confined to the primary ten-frame horizon in this evaluation.

S13.7 Fixed-task decision consequences and subject sensitivity

We decompose the existing content-preserving selection results while holding each history, future target, predictor and quality metric fixed. This retrospective audit reuses the ninety frozen pair/horizon/clip comparisons and all thirty native target-error records per model. It adds no forecasts. The primary pair remains SimVP–ConvLSTM, selected by first-frame raw C and evaluated by native ten-frame MSE; mean-ten probe readouts and PredRNN pairs remain secondary.

A decision-level decomposition. Let $s_r(i)$ be the model chosen by one fixed graph-only rule r on clip i, including a fixed fallback for unsupported inputs, score ties or abstentions. With $e_i(m)$ the native ten-frame error of model m, the policy risk is

$$R_r = \frac{1}{n}\sum_i e_i(s_r(i)), \qquad R_r - R_q = \frac{1}{n}\sum_{i:s_r(i)\neq s_q(i)} \big[e_i(s_r(i)) - e_i(s_q(i))\big].$$

The same rule is applied to every clip. This comparison does not choose reference weights or a preferred rule separately for each clip using target error. Each of the two fixed fallbacks is retained, with all-candidate and eligible-only denominators reported separately. Uniform random choice has exact expected error $(e_i(f) + e_i(g))/2$ and needs no sampled decisions.

Where the rule difference comes from. For the primary comparison, maximum entropy and maximum floor make identical choices. Minimum squared mass differs only on subject 22/walking: the former rules select SimVP and the latter selects ConvLSTM. The corresponding native errors are 4.9928×10^{-4} and 4.5280×10^{-4}. The all-thirty mean risk difference is $1.54939167 \times 10^{-6}$ under either common fallback. Removing subject 22 makes it zero; removing any other subject leaves $1.936739587 \times 10^{-6}$ on twenty-four clips. This identifies the full aggregate decision consequence of the observed fixed-rule disagreement.

Table 33: Primary policy risks on all thirty candidates. Probe decisions precede fallback; fixed and random baselines cover all candidates directly. Differences are relative to the same-cohort baseline. MSE is $\times 10^{-3}$.

Rule	Fallback	Decisions	MSE	vs ConvLSTM	vs random
Always SimVP	—	—	1.5072225	$+9.4762\%$	$+4.5238\%$
Always ConvLSTM	—	—	1.3767577	0.0000%	-4.5238%
Uniform random	—	—	1.4419901	$+4.7381\%$	0.0000%
Maximum entropy	SimVP	15	1.4956138	$+8.6330\%$	$+3.7187\%$
Maximum entropy	ConvLSTM	15	1.4318988	$+4.0051\%$	-0.6998%
Minimum squared mass	SimVP	15	1.4940644	$+8.5205\%$	$+3.6113\%$
Minimum squared mass	ConvLSTM	15	1.4303494	$+3.8926\%$	-0.8073%
Maximum floor	SimVP	15	1.4956138	$+8.6330\%$	$+3.7187\%$
Maximum floor	ConvLSTM	15	1.4318988	$+4.0051\%$	-0.6998%
Three-rule unanimity	SimVP	14	1.4956138	$+8.6330\%$	$+3.7187\%$
Three-rule unanimity	ConvLSTM	14	1.4303494	$+3.8926\%$	-0.8073%
Full-domain sign	SimVP	4	1.5132009	$+9.9105\%$	$+4.9384\%$
Full-domain sign	ConvLSTM	4	1.3740296	-0.1982%	-4.7130%

Fallback benefit and subject deletion. The full-domain policy uses a score ordering only when its entire admissible raw-gap interval excludes zero; otherwise it uses the specified fixed model. Its ConvLSTM fallback has a small all-candidate advantage, while the SimVP fallback increases risk relative to always-SimVP. Table 34 retains the original policy and removes one subject at a time, without refitting or rule selection. The sign reversal after removing subject 25 shows that the pooled ConvLSTM-fallback gain does not survive every deletion. These are descriptive sensitivity checks on five subject clusters, not independent replications or confidence intervals.

Table 34: Full-domain sign with ConvLSTM fallback after each subject deletion. Every row contains twenty-four original candidates. Positive values mean higher error than the stated baseline.

Omitted subject	vs always-ConvLSTM	vs uniform random
18	-0.1714%	-5.3287%
20	-0.3945%	-5.4429%
22	-0.2590%	-0.7776%
24	-0.2574%	-6.2253%
25	$+0.0800\%$	-5.4272%

The accompanying exact-arithmetic report retains 180 candidate/pair/horizon rows, 936 policy aggregates, 780 subject-deletion aggregates and 72 fixed-rule comparisons across both fallbacks and cohort scopes. It reconstructs choices from saved probe scores and checks 864 previously reported means exactly. The analysis separates finite reference-induced decision costs from evidence for a generally useful model-selection policy.

S14 Operational Meaning of the Reference Domain

The domain addresses a specified comparison problem: use eligible matched/reference pairs, equalize their aggregate history composition, and apply the same edge masses across clips when the named graph is common. These requirements motivate its constraints independently of whether a predictor's interval crosses zero. They do not establish how frequently practicing evaluators would choose a feasible weighting.

Balance as an exact cancellation requirement. Let $d = \sum_{(m,r)} w_{mr}(a_m - a_r)$, with a_s the endpoint-incidence vector of pair s. A pair score formed by adding arbitrary scalar scores g_j of its individual histories has matched-minus-reference contrast $g^\top d$. If $d = 0$, this contribution cancels for every g. Conversely, if it cancels for every g, taking g to be each coordinate vector gives $d_j = 0$ for every history. Endpoint balance is therefore necessary and sufficient for this cancellation property. It does not cancel general interactions between two histories. Prediction disagreement is such a pairwise score. Requiring fixed matched masses is an additional constraint; its infeasibility for construction B remains a change-of-estimand issue rather than a reason to silently relax it.

A support path instead of a preferred floor. For $0 \le \alpha \le 1$, the domains $\mathcal{P}_{\alpha t_*}$ are nonempty and nested. The base balanced domain is nonempty for these graphs and is a closed subset of the probability simplex, hence compact. The minimum edge mass is continuous, so its maximum is attained. A maximizing weighting establishes nonemptiness at every floor; the lower-bound inequalities establish nesting. Thus a strictly positive minimum or strictly negative maximum at zero floor establishes the same sign at every larger floor. Conversely, failure of strict one-sidedness at maximum floor implies failure at every smaller floor. A cell that first becomes one-sided along the path requires an additional minimum-mass assumption. A zero-floor certificate is particularly strong because it allows edges to be omitted; the positive-floor domains additionally require full support. The half-floor is retained for comparison with the earlier study, while the complete path identifies the role of that choice.

The original classification uses the strict tests $\ell > 0$ and $h < 0$, placing any zero-touching interval in the remaining category. The 17/20/37 coverage counts and the 17/20/3 path counts use this same convention. The existing two-artifact, two-construction, two-horizon grid on five KTH subjects yields 17 cells one-sided throughout, 20 requiring a positive floor, and three not strictly one-sided even at maximum floor (Table 17). These 40 cells are correlated finite comparisons. They are not a prevalence estimate, a cross-predictor ranking census, or evidence for selecting a particular floor. The critical brackets quantify where the additional assumption changes a direction, with no new forecasts.

Different checks answer different questions. The prospective B delete-one-clip ratios all remain below one: this checks deletion sensitivity of that reported ratio, not invariance to every weighting. In the five-subject grid, 36/40 half-floor categories agree after replacing pixel disagreement by structural disagreement: changing the readout changes the objective, not the feasible weight set. In the earlier 18-clip cohort, ConvLSTM B at mean ten is two-sided under clip-dependent weights but negative throughout the shared domain: imposing a common rule removes that ambiguity. Finally, the pure-permutation study changes the transition descriptor and admits 15/30 clips. These are distinct comparisons of data composition, objective, sharing convention and input eligibility. None is incorporated into a single LP interval.

Using the result in an evaluation report. An evaluator can specify the input eligibility rule, endpoint-composition requirement, sharing convention and readout, then distinguish a point result from its feasible range. The zero-floor range reports what those requirements imply without a minimum-mass assumption. The floor path states which extra minimum-mass restrictions establish a strict direction. A chosen central rule and its local radius describe behavior near that rule. This procedure makes the dependence on evaluation choices explicit; validating the prevalence of those choices among practitioners requires separate evidence.

S15 Connecting Order Response to Forecast Consequences

We reuse all fifteen admitted clips, all three frozen predictors and all 23 nonidentity anchored permutations from the content-preserving study. The original future remains the target. The outcome is the cost of this specified history perturbation, not a general temporal ability or architecture mechanism. The prior task-error values were already observed; the new response and alignment analysis is retrospective. Its plan, including symmetric consistency criteria, was fixed before these new readouts were computed. No forecasts, training or feature-model calls are added.

An exact error decomposition. For native prediction z, perturbed prediction z^π, target y and $\delta = z^\pi - z$, define $Q_{\mathrm{nat}} = \langle|\delta|\rangle$, $V = \langle\delta^2\rangle$ and $A = 2\langle(z-y)\delta\rangle$. With $N = \langle(z-y)^2\rangle$, excess target MSE satisfies

$$\Delta R = \langle(z^\pi - y)^2\rangle - N = V + A.$$

Angle brackets average pixels and the specified prediction horizon. Thus output movement can either reinforce or partly cancel native error. This elementary identity supplies an interpretation of the measurements, not a new theorem. Here ΔR is signed squared-error increase; it differs from the earlier absolute aggregate-error-change readout D. Likewise, Q_{nat} compares each perturbed forecast to the native forecast, whereas the main reference contrasts compare matched and reference pairs.

Aggregation and symmetric criteria. Mean ten is primary, the first prediction secondary. We average the 23 perturbations uniformly within each clip and then average clips equally. Every subject, pooled result and single-subject deletion is reported. For each of the three named model pairs, the signs of the pooled Q_{nat} and ΔR differences define concordance or discordance. A descriptive stability screen requires the same two nonzero signs after every subject deletion and in at least four of five subjects. Concordance and discordance use identical criteria; ties qualify for neither. Each subject counts once in this screen, including the two-clip subjects 18 and 24. This is not a significance test or an independent confirmation cohort.

Observed task connection. All three pooled primary comparisons are concordant. Only ConvLSTM/PredRNN meets the stability screen: PredRNN has both larger response and larger excess error in four subjects and after all five deletions. SimVP/PredRNN has three concordant subjects and passes all deletions; SimVP/ConvLSTM has two and fails the deletion condition. No primary pair meets the discordance screen. This connects a named response comparison to a measured task consequence under the same perturbation distribution. It neither validates all response rankings nor supplies evidence that practitioners use the endpoint-balanced weight domain.

Squared movement and excess error have the same pooled model ordering. Although the pooled cosine magnitudes are below 0.011, alignment offsets 26.95% of squared movement for SimVP and 11.30% for PredRNN; a small cosine therefore does not imply a negligible contribution relative to movement. ConvLSTM has almost no pooled cancellation. Subject 24 provides a local counterexample: PredRNN's negative alignment more than offsets its movement and yields a negative excess MSE. This exception is retained, not promoted to a stable reversal. We report $A/(2\sqrt{NV})$ as the cosine over the weighted concatenation of error and perturbation fields, rather than an average of per-example cosines. Ratios A/V and $V/\Delta R$ are descriptive, signed and undefined at zero denominators; they enter no stability criterion.

Table 35: Pooled mean-ten response and task-error decomposition, fifteen clips and 23 fixed perturbations per clip. Q_{nat} is scaled by 10^3; $V, A, \Delta R$ by 10^6.

Model	Q_{nat}	V	A	ΔR	$\Delta R/N$
SimVP	0.9361	7.1537	-1.9277	5.2260	0.3399%
ConvLSTM	0.5648	4.7469	0.0042	4.7511	0.3383%
PredRNN	1.5280	46.4797	-5.2512	41.2285	3.0452%

Complete primary subject and deletion results. Table entries are independently rounded; the identity holds on unrounded values. The same units apply below. All native MSE values N, including subject and deletion scopes, are provided as `native_mse` in the companion `RESULTS.json`.

Pooled and deleted-subject results reuse the same clips. The first-frame secondary grid and all pair signs accompany the saved results file.

Scope	Model	Q_{nat}	V	A	ΔR	Cosine
pooled	SimVP	0.9361	7.1537	-1.9277	5.2260	-0.0092
pooled	ConvLSTM	0.5648	4.7469	0.0042	4.7511	0.0000
pooled	PredRNN	1.5280	46.4797	-5.2512	41.2285	-0.0105
subject-18	SimVP	1.3095	8.9075	3.3029	12.2104	0.0142
subject-18	ConvLSTM	0.7170	3.8621	11.0446	14.9066	0.0699
subject-18	PredRNN	2.6937	86.6709	179.4406	266.1115	0.2551
subject-20	SimVP	0.8742	6.4159	2.7835	9.1994	0.0144
subject-20	ConvLSTM	0.5875	5.6772	-7.2371	-1.5599	-0.0404
subject-20	PredRNN	1.3496	36.2483	15.1747	51.4229	0.0328
subject-22	SimVP	0.6078	1.9712	0.1180	2.0893	0.0010
subject-22	ConvLSTM	0.2647	0.4038	-0.3934	0.0104	-0.0097
subject-22	PredRNN	0.8775	7.1860	11.7952	18.9813	0.0703
subject-24	SimVP	1.2496	16.8768	-14.9469	1.9299	-0.0360
subject-24	ConvLSTM	1.0818	14.1735	3.4082	17.5817	0.0095
subject-24	PredRNN	2.7534	137.4590	-261.6404	-124.1814	-0.2287
subject-25	SimVP	0.9095	5.9147	-6.6330	-0.7183	-0.0451
subject-25	ConvLSTM	0.3808	1.8450	2.8410	4.6860	0.0326
subject-25	PredRNN	0.8820	15.3789	-8.5425	6.8364	-0.0386
delete-18	SimVP	0.8786	6.8839	-2.7324	4.1515	-0.0133
delete-18	ConvLSTM	0.5414	4.8830	-1.6943	3.1887	-0.0104
delete-18	PredRNN	1.3487	40.2965	-33.6653	6.6311	-0.0724
delete-20	SimVP	0.9670	7.5226	-4.2833	3.2394	-0.0197
delete-20	ConvLSTM	0.5534	4.2817	3.6248	7.9066	0.0234
delete-20	PredRNN	1.6173	51.5955	-15.4642	36.1313	-0.0299
delete-22	SimVP	1.0181	8.4493	-2.4391	6.0102	-0.0108
delete-22	ConvLSTM	0.6398	5.8327	0.1036	5.9363	0.0006
delete-22	PredRNN	1.6907	56.3032	-9.5128	46.7903	-0.0167
delete-24	SimVP	0.8878	5.6578	0.0753	5.7331	0.0004
delete-24	ConvLSTM	0.4852	3.2967	-0.5195	2.7771	-0.0040
delete-24	PredRNN	1.3395	32.4829	34.1933	66.6762	0.0867
delete-25	SimVP	0.9427	7.4635	-0.7514	6.7121	-0.0033
delete-25	ConvLSTM	0.6108	5.4724	-0.7050	4.7673	-0.0039
delete-25	PredRNN	1.6896	54.2549	-4.4284	49.8266	-0.0078

Independent verification. Remote CPU computation reads unmodified float32 outputs with float64 accumulation and no clipping. All 2,070 horizon-specific rows reconcile their native and perturbed MSEs against the earlier task analysis (relative tolerance 10^{-12}, absolute 10^{-15}). A separate frame-wise dot-product implementation checks the decomposition, all 90 clip aggregates, 66 scope/model aggregates, 66 pair comparisons and three stability screens. The largest decomposition residual is 1.63×10^{-18}; independent readouts differ by at most 2.17×10^{-18}. These checks verify arithmetic, not population generalization.

S16 Prospective Confirmation on Later Video Windows

The analysis in Appendix S15 selected PredRNN versus ConvLSTM as the sole mean-ten comparison meeting its symmetric stability screen. Before decoding new windows or acquiring forecasts, we fixed this pair, its positive PredRNN-minus-ConvLSTM direction, the perturbations and the confirmation criterion. Confirmation requires positive pooled differences in both Q_{nat} and signed excess MSE ΔR, the same two positive differences after each of five subject deletions, and both positive in at least four of five subjects. Ties fail; the first-frame horizon is secondary. No alternate pair or pooled combination of old and new windows can replace this criterion.

Acquisition. We retain the fifteen previously admitted d2 videos and take frames 21–40 of their original annotated action intervals: frames 21–30 are history and frames 31–40 are the unperturbed target. These windows do not overlap the original frames 1–20. All fifteen predefined windows are available; each new history has ten distinct frames. We apply all 24 anchored permutations of positions 2–5, keeping the first and final five frames fixed. The identity supplies the native forecast and the other 23 receive equal weight. No graph-admission filter is reapplied to these windows: the task comparison concerns this fixed perturbation distribution. We average perturbations within each clip and then weight clips equally.

The checkpoints and inference adapters are unchanged. Two engineering forwards reproduced the old cached predictions byte for byte before the 720 scientific forwards (360 per model). Every input and output has a hash receipt. Both models finished all inputs; an independent framewise dot-product recount verified all 1,380 clip/permutation/model/horizon rows and the aggregate and confirmation calculations. No training or model selection was performed. The new windows come from the same videos and five subjects, so nonoverlap does not establish statistical independence or new-subject generalization.

Primary outcome. The new-window result meets the fixed confirmation criterion. Both pooled differences are positive, 5/5 subjects have both positive differences, and all subject deletions retain both positive differences. PredRNN and ConvLSTM have pooled responses 2.970 and 1.389 ($\times 10^{-3}$), respectively; their signed excess MSEs are 159.01 and 0.57 ($\times 10^{-6}$). The ratios of pooled excess to native MSE are 16.65% for PredRNN and 0.049% for ConvLSTM. This prospectively reproduces the selected response–task-error relation on later windows within the same videos.

Complete directional results. Table 37 retains every subject and deletion at both horizons. All differences are PredRNN minus ConvLSTM. Negative excess MSE is retained and denotes improvement relative to the native forecast, not failure of the measurement.

Table 37: Fixed-pair confirmation. ΔQ is in 10^{-3} units; $\Delta\Delta R$ is in 10^{-6} units. Positive values in both columns are the prespecified direction.

	Mean ten (primary)		First frame (secondary)	
Scope	ΔQ	$\Delta\Delta R$	ΔQ	$\Delta\Delta R$
pooled	1.581	158.44	0.491	17.95
subject 18	0.808	57.19	0.241	4.86
subject 20	1.821	215.76	0.591	36.52
subject 22	1.031	30.61	0.355	5.95
subject 24	3.043	434.00	0.662	7.80
subject 25	1.269	74.55	0.514	14.49
delete 18	1.699	174.02	0.529	19.96
delete 20	1.460	129.79	0.441	8.66
delete 22	1.718	190.40	0.525	20.95
delete 24	1.355	116.05	0.465	19.51
delete 25	1.659	179.42	0.485	18.81

Movement and error alignment. The decomposition $\Delta R = V + A$ uses $V = \langle \delta^2 \rangle$ and $A = 2\langle (z - y)\delta \rangle$. Table 38 reports all pooled components rather than inferring task harm from response magnitude alone. At mean ten, ConvLSTM's negative alignment nearly offsets its squared movement ($45.015 - 44.449 = 0.566$, $\times 10^{-6}$), whereas PredRNN retains a larger error increase ($172.394 - 13.385 = 159.009$). Response alone does not encode this cancellation.

S17 Supporting Checks and Decision Boundary

The earlier controlled RGB study shows why matching must start from actual inputs. A repeated prefix collapses 120 nominal permutations into five histories; a matched early/late placement then gives instance-dependent responses across six frozen checkpoints. The initial public slice and shorter-loop follow-up establish the constructions used in the prospective grid. They reuse inspected clips and serve as development evidence, rather than additional independent replications (Appendices S27.1–S27.2).

Table 38: Pooled new-window task quantities, all except Q in 10^{-6} units; Q in 10^{-3}.

Horizon	Model	Q	V	A	ΔR	Native MSE
mean10	PredRNN	2.970	172.39	-13.39	159.01	954.80
mean10	ConvLSTM	1.389	45.01	-44.45	0.57	1161.52
first	PredRNN	1.120	18.96	0.33	19.30	172.33
first	ConvLSTM	0.629	4.37	-3.02	1.35	229.78

A separate, prespecified native-history selection screen asks a different question: can transition magnitude or its temporal centroid choose between the two artifacts? On 128 development clips, both four-cell rules choose SimVP everywhere and produce zero improvement over the global choice. The predeclared development stop leaves confirmation unevaluated (Appendix S27.3). This result limits a proposed selection use; it does not test selection conditioned on our constructed block-exchange responses.

Table 39: Initial and prospective first-frame SimVP Q ratios, unmasked. The final row changes matched and reference weights; ratios are not forecast accuracy.

Study/family	Comparison	Ratio	Status
Initial/A	All references	1.1466	Initial observation
Initial/A	Exact distance	1.0633	Post hoc
Follow-up/A	Exact distance	0.9996	Positive criterion not met
Follow-up/B	Exact distance	0.8978	Positive criterion not met
Follow-up/B	Balanced witness	0.9870	Post-hoc new weights

S18 Related Work

Transition-preserving sequence controls. Constrained surrogate methods preserve selected observed properties while testing for structure beyond them; their interpretation depends on the null and constraint design (Schreiber & Schmitz, 2000). Our six deterministic arrangements do not implement that statistical testing framework. Pethel and Hahs construct surrogate sequences with fixed transition counts and endpoints for Markov-order testing (Pethel & Hahs, 2014). Transition-preserving rearrangement is therefore an established principle. Our deterministic closed-block pairs apply that principle to pixel-prediction responses with fixed factual targets; they neither implement a uniform surrogate sampler nor supply a Markov-order hypothesis test. The contribution rests on the measured prediction behavior and its accounting, not a new transition-count invariance.

Temporal probes. Shuffle and Learn uses temporal-order verification for video representation learning (Misra et al., 2016); its sampled tuples retain endpoints while replacing an intermediate frame. Atemporal probes study image-level contributions to video-language performance (Buch et al., 2022), while TempCompass separates shared static content from temporal properties (Liu et al., 2024). A recent preprint intervenes on information channels in video-VLM evaluation (Rahman, 2026). Our construction instead preserves the complete directed adjacent-frame-pair multiset of two prediction inputs. It tests the temporal assignment left unspecified by that match, without identifying the predictor's internal channels.

Prediction and evaluation. Moving MNIST and OpenSTL provide the prediction setting and released artifacts (Srivastava et al., 2015; Tan et al., 2023; Li et al., 2023). ContextVP studies access to past context and uses motion-based regional scoring (Byeon et al., 2018); our target-intensity regions answer a different localization question and are not motion masks. Evaluation criteria for generative models need not agree (Theis et al., 2016). FVD measures video distributions (Unterthiner et al., 2018), and its content sensitivity has been studied under temporal corruption (Ge et al., 2024); VBench separates quality dimensions (Huang et al., 2024). Here the unit is a paired pixel reconstruction error. Exact input matching and pointwise output accounting expose distinct losses of information that a scalar summary alone cannot resolve. The identities used for that accounting are elementary; their role is to interpret the complete observed prediction grid.

Discriminative diagnostics versus forecast comparisons. Hwang et al. (2024) connect action-recognition robustness to shuffled frames with reduced motion reliance and residual motion monotonicity. TemporalLens (Tomotaki-Dawoud et al., 2026) probes video detectors through structured perturbations, occlusion, shuffling and redundancy. Alrasheed et al. (2026) compare video representations across robustness axes including temporal direction. These studies motivate distinguishing task performance from temporal dependence. Our question concerns pixel forecasts and the reference choices within a controlled comparison; neither shuffle robustness in recognition nor semantic directionality establishes the meaning of a forecast contrast. Full reversal preserves frame content but generally changes boundary frames, the recent suffix and directed transitions. It is consequently a complementary diagnostic, not an exchangeable matched control under our construction. A native/reversed pair alone supplies no matched/reference family over which to infer LP ambiguity.

S19 Complete Prospective Six-Permutation Results

The fixed grid has 896 study forecasts and 2 separate engineering smoke forecasts, with zero training. Indices 320–351 and all comparisons were frozen before outputs; future targets were extracted only after the study forecasts were complete and archived. Input content is disjoint from 320 recorded prior native histories. The new targets come from the same fixed public Moving MNIST file through explicit byte ranges; no confirmation target from the stopped selection study was opened.

Figure 11 displays every primary unmasked pair for each artifact, alongside mean input distances. Complete Q, L, D and error tables remain in the CSV supplement. Table 40 retains both arms and both horizons. The accompanying `supplement/r13` directory contains all 1,792 condition rows, 3,840 pair rows, 19,200 per-step rows, 1,920 temporal-cancellation rows, 512 within-clip group rows and 256 group-difference rows. The 896 forecast identities map to immutable raw outputs in the experiment archive. JSON summaries include means, medians, extrema, signs and all leave-one-clip mean ranges. No pair or zero response is omitted.

Actual matched-pair counts are three for every V clip and 31 M clips, but seven for clip 324/M. At that exception the classes are $\{P_{012}, P_{021}, P_{120}, P_{210}\}$ and $\{P_{102}, P_{201}\}$. The additional exact-history identities are $P_{012} = P_{021}$ and $P_{120} = P_{210}$. They are recorded before inference and remain in all named-pair summaries. Group weights use these actual classes; named-pair summaries always use all 32 clips. Numerically negligible roundoff in an algebraically balanced pair-error difference is not interpreted as a scientific direction.

Table 40: Means of within-clip group averages using actual pixel-directed-edge equivalence. Most clip/arms have 3 matched and 12 unmatched pairs; clip324/M has 7 and 8. All 32 clips receive equal weight. These are descriptive groups, not randomized treatments.

Artifact	Arm	Readout	Pairs	Q	L	D
simvp	V	first	matched	0.00389	0.00366	0.00203
simvp	V	first	unmatched	0.00339	0.00320	0.00175
simvp	V	mean10	matched	0.00883	0.00824	0.00351
simvp	V	mean10	unmatched	0.00744	0.00700	0.00291
simvp	M	first	matched	0.00245	0.00238	0.00099
simvp	M	first	unmatched	0.00229	0.00223	0.00092
simvp	M	mean10	matched	0.00604	0.00584	0.00220
simvp	M	mean10	unmatched	0.00542	0.00525	0.00193
convlstm	V	first	matched	0.10782	0.08761	0.01954
convlstm	V	first	unmatched	0.10805	0.08795	0.02014
convlstm	V	mean10	matched	0.14372	0.12146	0.04006
convlstm	V	mean10	unmatched	0.14417	0.12231	0.04608
convlstm	M	first	matched	0.09496	0.07682	0.01619
convlstm	M	first	unmatched	0.09496	0.07707	0.01943
convlstm	M	mean10	matched	0.14399	0.11872	0.03607
convlstm	M	mean10	unmatched	0.14387	0.11843	0.03649

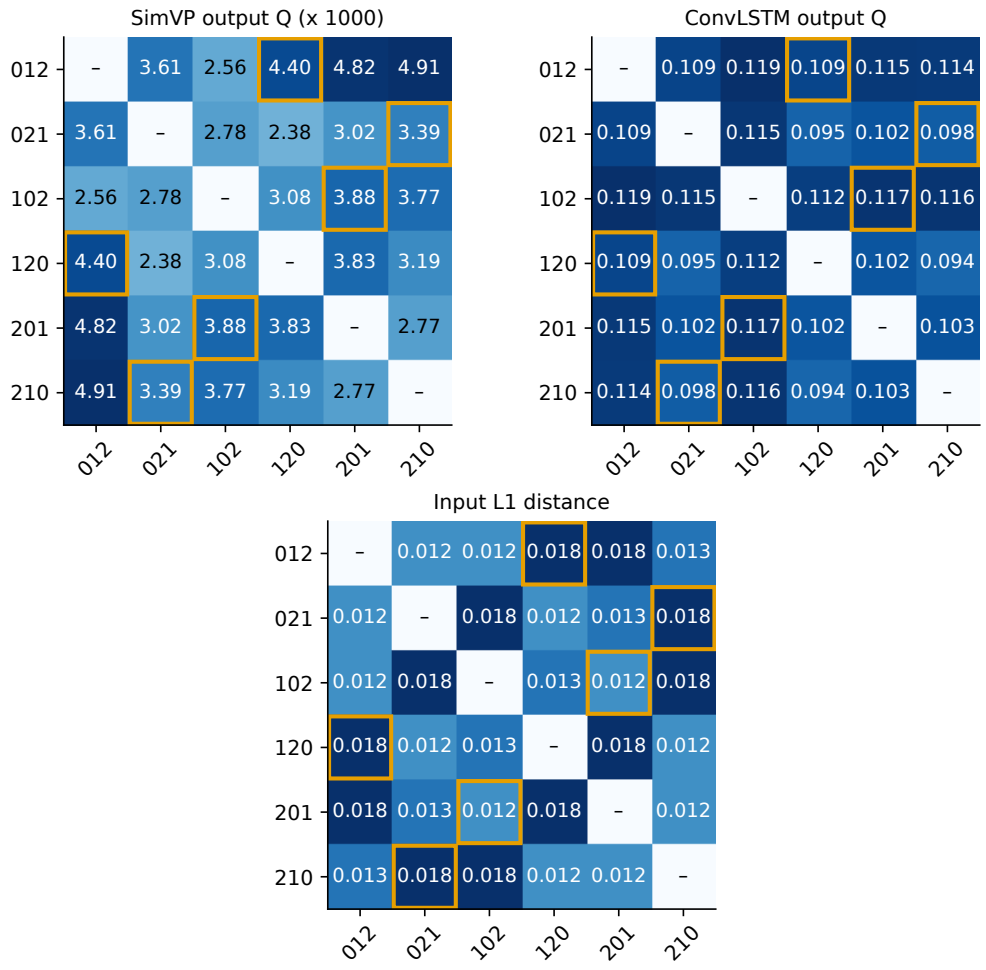

Figure 11: All fifteen unordered pairs in arm V, averaged over the same 32 clips; matrices are symmetric. Orange outlines mark directed-transition matches. SimVP Q is multiplied by 1,000 for legibility; ConvLSTM Q and normalized input L1 are unscaled. Input-distance matching is done exactly within each clip, not on these displayed means.

Table 41: Fixed-pair replication on new clips 320–351. Pair is the mean error of P_{102} and P_{201}; ratio is pair/native means. First is primary; mean10 is secondary. All entries retain 32 clips.

Artifact	Arm	Readout	Native	Pair	Ratio	Q	L	D
simvp	V	first	0.00874	0.01038	1.19	0.00388	0.00370	0.00211
simvp	V	mean10	0.01188	0.01702	1.43	0.00777	0.00733	0.00273
simvp	M	first	0.03273	0.03394	1.04	0.00225	0.00218	0.00093
simvp	M	mean10	0.03871	0.04304	1.11	0.00540	0.00522	0.00204
convlstm	V	first	0.01366	0.09193	6.73	0.11662	0.09394	0.02196
convlstm	V	mean10	0.01719	0.13389	7.79	0.16193	0.13498	0.04350
convlstm	M	first	0.03479	0.09246	2.66	0.10199	0.08366	0.01703
convlstm	M	mean10	0.04232	0.12753	3.01	0.14008	0.11676	0.03066

S19.1 Initial fixed-pair error checks

The short matched-pair result repeats on new clips. For SimVP in arm V, pair-average error is 0.01038 versus native 0.00874, a ratio of 1.19. Output disagreement, pixelwise error change and absolute aggregate error difference average $Q = 0.003885$, $L = 0.003700$ and $D = 0.002106$. The median D is 0.001947, with 16 positive and 16 negative signed differences. The signed mean,

-0.000119, is much smaller than the mean absolute response. On the earlier exploratory slice, the same pair had error 0.00976, native ratio 1.25 and mean $D = 0.00145$ (Appendix S27.1). These are different clip sets, so this is a side-by-side replication description rather than a paired improvement claim.

For ConvLSTM, the new unmasked pair error is 0.09193 versus native 0.01366, a ratio of 6.73. Its $Q/L/D$ means are 0.11662/0.09394/0.02196, with 9 positive and 23 negative differences. The earlier slice had native ratio 6.82 and mean $D = 0.01621$. The artifact continues to respond at a much higher relative error scale than SimVP. A large response is therefore not evidence of good prediction. Masked and ten-frame results differ quantitatively (Table 41); each ratio uses its own arm and horizon's native error.

S20 Post-Review Distance, Step and Range Analyses

These three analyses were planned after review of the original grid, before their new summaries were computed. They reuse the 896 archived forecasts and add no model execution. Primary experiment summaries remain unchanged and separately archived. No p-values or population confidence intervals are introduced.

Distance definitions and weighting. For uint8 histories x^A, x^B, define $s(A,B) = \sum_{t,p} |x^B_{t,p} - x^A_{t,p}|$ and $m_p(A,B) = \sum_t |x^B_{t,p} - x^A_{t,p}|$. Scalar matching requires exactly equal integer s; spatial matching additionally requires m_p equality for every one of the 4,096 pixels. There is no tolerance, nearest-neighbor choice or output-dependent selection. For each actual matched pair, all qualifying actual unmatched pairs receive equal weight. Covered matched pairs then receive equal weight within their clip, and all covered clips receive equal weight. A zero denominator is undefined, not stabilized. The endpoint-balance proof for original exhaustive groups need not hold under this reweighting.

In V, both rules cover all 96 matched comparisons. In M, both cover 97/100: the three identical-input pairs P_{012}/P_{021}, P_{102}/P_{201} and P_{120}/P_{210} in clip324 have no zero-distance unmatched reference. Their exact-zero responses remain in original summaries and are explicitly marked uncovered here. The four other actual matched pairs in that clip each use all eight unmatched references. All 32 clips still contribute, with the stated covered-pair weighting. Covered masked means are therefore a different estimand from the original full matched group.

The scalar and spatial reference sets differ once: clip340/V, pair P_{102}/P_{201}. Its integer distance is 134,408; scalar matching admits five references, whereas spatial matching admits two. All five are retained under the scalar rule. For ordinary matched pairs, P_{102}/P_{201} changes two symbolic slots; the other two matched pairs change three. Thus not all matched pairs are three-slot cyclic rotations. Complete distances, reference edges, missing coverage, per-pair and per-clip summaries are supplied under `supplement/r13/post_review_distance`.

Table 42: Post-review exact-distance references. Ratios use equal clip weights after averaging all references within each matched pair and covered matched pairs within each clip. S: scalar L1; P: pixelwise sum map. Coverage is 96/96 in V and 97/100 in M; all 32 clips contribute.

Artifact	Arm	Horizon	Rule	Q ratio	L ratio	D ratio
simvp	V	first	S	1.0633	1.0641	1.0788
simvp	V	mean10	S	1.1157	1.1097	1.1109
simvp	M	first	S	0.9967	0.9987	0.9995
simvp	M	mean10	S	1.0434	1.0435	1.0535
convlstm	V	first	S	0.9919	0.9906	0.9439
convlstm	V	mean10	S	0.9928	0.9899	0.8763
convlstm	M	first	S	0.9962	0.9908	0.8152
convlstm	M	mean10	S	1.0032	1.0077	1.0584
simvp	V	first	P	1.0613	1.0621	1.0749
simvp	V	mean10	P	1.1141	1.1081	1.1053
simvp	M	first	P	0.9967	0.9987	0.9995
simvp	M	mean10	P	1.0434	1.0435	1.0535
convlstm	V	first	P	0.9940	0.9927	0.9480
convlstm	V	mean10	P	0.9952	0.9913	0.8718
convlstm	M	first	P	0.9962	0.9908	0.8152
convlstm	M	mean10	P	1.0032	1.0077	1.0584

Every future step. Stepwise summaries use the original actual matched and unmatched groups, with pair means inside clips followed by equal clip weighting. They are separate from the distance-controlled summaries. For unmasked SimVP, the ten Q ratios range from 1.147 to 1.207, L from 1.143 to 1.196 and D from 1.117 to 1.242. Masked Q and L fall below one at step three. ConvLSTM's unmasked Q ratios range from 0.992 to 1.001, whereas all ten D ratios are below one (0.837–0.970). These response curves do not establish an integration mechanism, a monotonic trend, or improvement over native prediction.

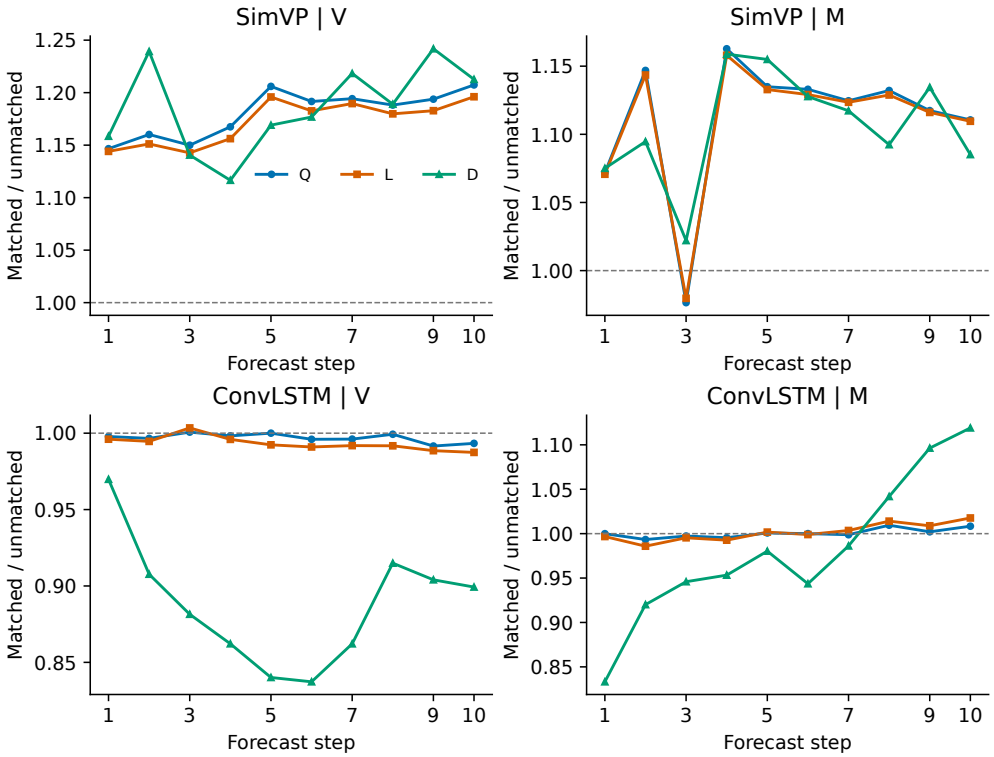

Figure 12: Post-review group response ratios at all ten future steps. Dashed lines mark one. Axes have panel-specific limits; values are ratios of equally weighted clip means. All steps and both arms are retained, including reversals. These are original exhaustive groups, without input-distance reweighting.

Exact output-range accounting. Since clipping and its signed residual are monotone, the two corresponding pair differences have the same sign; their magnitudes sum to the raw pair difference. We verify $Q = Q_c + Q_e$ pointwise over every saved pair, with maximum residual zero in the float64 computation. This signed-residual discrepancy $Q_e = \langle|(z_B - c_B) - (z_A - c_A)|\rangle$ is distinct from the difference of scalar excursion errors. The clipped error responses L_c, D_c are scored against the unchanged factual targets. All 15 pairs are evaluated for both arms and both horizons. Original actual group weights, including clip324/M, are retained.

For the first unmasked frame, ConvLSTM's matched and unmatched excursion shares are 25.922% and 25.919%. Nearly three quarters of its disagreement therefore survives clipping in both groups. Their Q_c ratio is 0.9978, so the near-flat raw comparison does not disappear when excursion is removed. SimVP's corresponding shares are 1.761% and 1.763%, with Q_c ratio 1.1466. These are arithmetic decompositions of outputs; no hidden activation or saturation mechanism was measured. Table 43 retains all groups and horizons.

Table 43: Post-review output-range accounting for original complete groups. Excursion share is the ratio of mean Q_e to mean Q, not a mean of per-clip fractions. M324 retains its original seven/eight grouping.

Artifact	Arm	Horizon	Group	Q	Q_c	Q_e/Q (%)	D_c/Q_c
simvp	V	first	mat.	0.00389	0.00382	1.76	0.533
simvp	V	first	unmat.	0.00339	0.00333	1.76	0.528
simvp	V	mean10	mat.	0.00883	0.00874	0.97	0.402
simvp	V	mean10	unmat.	0.00744	0.00736	1.01	0.397
simvp	M	first	mat.	0.00245	0.00236	3.83	0.426
simvp	M	first	unmat.	0.00229	0.00220	3.81	0.423
simvp	M	mean10	mat.	0.00604	0.00585	3.18	0.377
simvp	M	mean10	unmat.	0.00542	0.00524	3.26	0.368
convlstm	V	first	mat.	0.10782	0.07987	25.92	0.178
convlstm	V	first	unmat.	0.10805	0.08005	25.92	0.182
convlstm	V	mean10	mat.	0.14372	0.11109	22.70	0.253
convlstm	V	mean10	unmat.	0.14417	0.11154	22.63	0.293
convlstm	M	first	mat.	0.09496	0.06609	30.40	0.162
convlstm	M	first	unmat.	0.09496	0.06613	30.36	0.198
convlstm	M	mean10	mat.	0.14399	0.10566	26.62	0.234
convlstm	M	mean10	unmat.	0.14387	0.10562	26.58	0.243

Simple suffix-only references. All six histories share their last five frames. Therefore any deterministic predictor using only that suffix, including copy-last and last-two-frame linear extrapolation with a fixed clipping convention, returns the same prediction for every history. Its Q, L, D responses are exactly zero by identical inputs. This is an analytic reference, not an additional model execution or evidence of good prediction; its absolute accuracy is not assessed on the new slice.

S21 Prospective Follow-up: Full Readouts and Reference Weights

The frozen plan, both exact-distance rules and all readouts preceded the 2,176 forecasts. Raw outputs were exported and checked before target frames 10–19 were acquired. The fixed input graph contains 3,840 unordered pair instances across clip/arm/family and 9,728 matched-reference edges counting both rules; no pixel degeneracy adds a class on this slice. The two rules' reference sets are identical. Every future step and all reversed comparisons are retained. Pair multiplicity does not enlarge the 32-clip unit of description.

For any pair, write $c = \mathrm{clip}(z, 0, 1)$ and signed excursion $e = z - c$. Pointwise, $|z_B - z_A| = |c_B - c_A| + |e_B - e_A|$ and

$$(|z_B - y| - |z_A - y|) = (|c_B - y| - |c_A - y|) + (|e_B| - |e_A|).$$

The latter identity is for signed loss; L and D are not additive across these terms. In the distance-weighted unmasked first-frame matched groups, signed-excursion disagreement accounts for 1.78%/1.21% of raw Q for SimVP A/B and 24.95%/26.92% for ConvLSTM A/B. These are ratios of weighted means, not averages of pixelwise ratios.

Reference weights give equal mass to each covered matched pair, then each clip. Within B, $(1, 1, 1)$ has four matched pairs with twelve references each; $(2, 2, 0)$ has six with two each; $(0, 0, 2)$ has two with three each. Every endpoint receives half of its incident pair mass; endpoint weights and every reused reference are saved explicitly. The full condition/pair tables include absolute errors and native ratios alongside response magnitudes. No formal equivalence test, population interval, accuracy gain or decision-utility claim is inferred from these ratios.

Independent CPU implementations recounted raw pair metrics, clip/reference means, coefficient strata, E/L directions, endpoint weights and common-clip comparisons. Some auxiliary pair-average/native-error differences are algebraically zero; floating summation order can change their last-bit signs. Such auxiliary sign counts are retained as arithmetic records and are not scientific direction evidence. The primary Q signs do not have this ambiguity. Complete tables and verification receipts accompany the manuscript.

Table 44: Complete prospective first/mean10 exact-distance response ratios. Both rules coincide on this slice. Delete-one ranges are finite sensitivity summaries for Q, not confidence intervals. Mean10 D takes the absolute value after averaging signed error differences over all ten steps.

Artifact	Arm	Family	Horizon	Q	L	D	Delete-one Q
SimVP	V	A	first	0.9996	0.9979	1.0339	0.9813–1.0125
SimVP	V	A	mean10	1.0209	1.0207	1.0151	1.0033–1.0376
SimVP	V	B	first	0.8978	0.9044	0.9523	0.8835–0.9076
SimVP	V	B	mean10	0.9187	0.9223	0.9161	0.9056–0.9282
SimVP	M	A	first	1.0185	1.0187	1.0438	0.9996–1.0274
SimVP	M	A	mean10	1.0086	1.0076	1.0283	0.9966–1.0231
SimVP	M	B	first	0.8593	0.8619	0.8552	0.8529–0.8794
SimVP	M	B	mean10	0.9149	0.9161	0.8647	0.9103–0.9244
ConvLSTM	V	A	first	1.0006	1.0017	0.9778	0.9985–1.0021
ConvLSTM	V	A	mean10	0.9864	0.9758	0.8410	0.9823–0.9924
ConvLSTM	V	B	first	1.0047	1.0020	1.0508	1.0022–1.0100
ConvLSTM	V	B	mean10	1.0064	0.9996	0.9463	1.0007–1.0144
ConvLSTM	M	A	first	1.0108	1.0128	0.8441	1.0078–1.0132
ConvLSTM	M	A	mean10	0.9947	0.9913	0.9367	0.9886–1.0030
ConvLSTM	M	B	first	1.0224	1.0211	0.9668	1.0115–1.0295
ConvLSTM	M	B	mean10	1.0058	1.0042	0.8945	0.9936–1.0148

Table 45: Every matched coefficient stratum, SimVP/V/first. Coefficients (n_{01}, n_{02}, n_{12}) count aligned unequal source-index pairs; actual pixel distances, not these coefficients, select references.

Family	Coefficient	Pairs/clip	Matched Q	Reference Q	Ratio
A	(0,0,2)	1	0.003443	0.003606	0.9548
A	(1,1,1)	2	0.004125	0.004046	1.0196
B	(0,0,2)	2	0.003915	0.004042	0.9687
B	(1,1,1)	4	0.004710	0.004588	1.0265
B	(2,2,0)	6	0.004612	0.005682	0.8118

Table 46: All six named E/L contrasts, SimVP/V/first. Signed loss is always L minus E, even where aliases change canonical names. Other arms, artifacts and horizons remain in the complete machine-readable tables. These six contrasts are distinct from the twelve-pair matched estimator and are not independent sample units.

(a, b, c)	Q	L	D	Signed loss
012	0.005216	0.004946	0.002290	0.000978
021	0.004661	0.004496	0.002300	0.001372
102	0.003453	0.003324	0.001586	0.000593
120	0.005051	0.004757	0.002482	0.001137
201	0.004644	0.004450	0.002384	-0.001309
210	0.003912	0.003767	0.001690	-0.000645

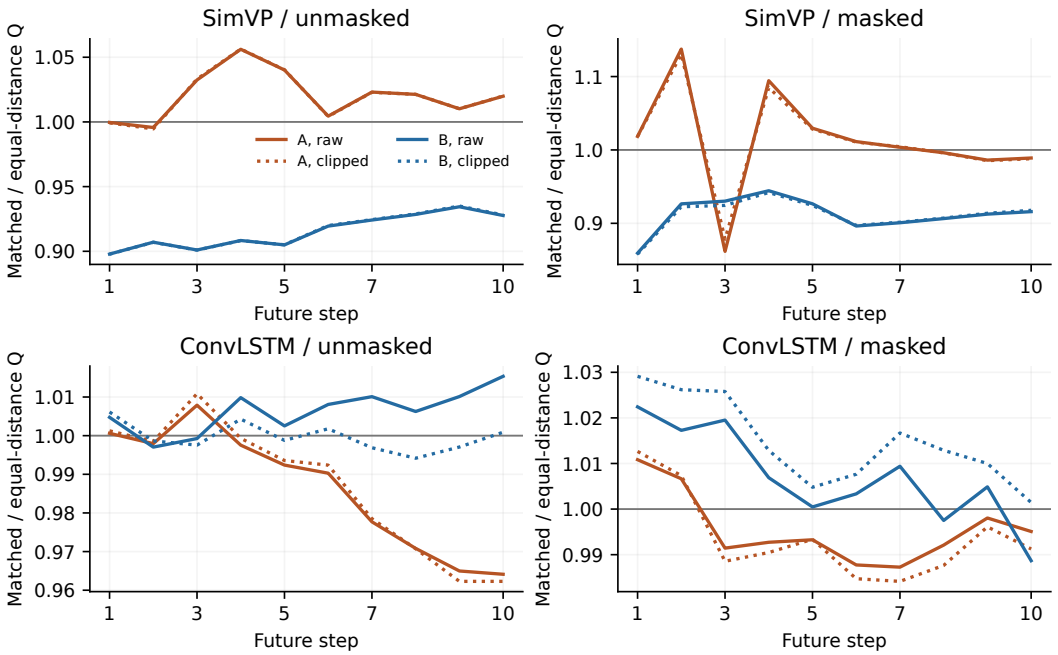

Figure 13: Prospectively fixed exact-distance prediction-disagreement ratios at all ten future steps. Color identifies family; solid lines use raw outputs and dotted lines clip outputs to $[0, 1]$. Each panel retains its own vertical scale for legibility. Both exact-distance rules coincide; neither future steps nor rules are independent replications. The first step is the primary horizon.

S22 Reference Accounting and Balance Feasibility

This analysis reuses the follow-up's 32 clips, both artifacts and V/M arms. First, mean10 and each of ten steps are retained for raw/clipped Q, L, D. No new prediction, target acquisition or training is used. The initial and follow-up studies still total 3,072 study forecasts. All results here are post hoc; prospective non-confirmation in Section S5 is unchanged.

Reference-mean identity. Let p_s and q_s be original and exact-reference masses of coefficient stratum s, and μ_s its original uniform reference mean. With exact pair weights w_r, the reference change is

$$R_{\text{exact}} - R_{\text{original}} = \sum_s (q_s - p_s)\mu_s + \sum_r w_r \big(Q_r - \mu_{s(r)}\big).$$

All original strata remain in the sum, including those assigned zero exact mass. The second term vanishes here because all allowed references in each realized stratum are uniformly weighted. The same accounting holds for each saved readout. It is arithmetic, not a causal decomposition. The scalar and pixelwise-map rules coincide with coefficient support on all 128 clip/arm/family graphs; symbolic equality alone would not exclude additional pixel coincidences on other data.

Linear balance problem. For a matched pair m and reference pair r, let a_m, a_r be endpoint incidence vectors with two ones. Joint masses $x_{mr} \geq 0$ are restricted to exact-distance support. Given matched weights u_m, normalized balance requires

$$\sum_r x_{mr} = u_m, \qquad \sum_m u_m = 1, \qquad \tfrac{1}{2}\sum_{m,r} x_{mr} a_r = \tfrac{1}{2}\sum_m u_m a_m.$$

If p^M, p^R denote the resulting endpoint marginals, then for any single-history score f, the difference of weighted pair-average scores is $\sum_v (p_v^M - p_v^R) f_v$. Balance makes this difference zero; it does not fix paired disagreement or establish a temporal mechanism.

B's equal-matched infeasibility follows from the class-mass contradiction in Section S6. The complete twelve-history symbolic family adds prefixes 010020 and 020010 before suffix 6789. Its classes have

Table 47: First-frame reference-mean accounting for raw Q. Between and within terms sum to exact minus original reference mean; matched means are unchanged in this comparison. Within-stratum weights remain uniform on this realized graph, so the within term is zero up to numerical roundoff.

Artifact	Arm	Family	Original ref.	Exact ref.	Between	Within
SimVP	V	A	0.003676	0.003899	0.000223	0.000000
SimVP	V	B	0.004249	0.005044	0.000795	0.000000
SimVP	M	A	0.002379	0.002507	0.000129	0.000000
SimVP	M	B	0.002626	0.003207	0.000580	0.000000
ConvLSTM	V	A	0.101750	0.101964	0.000214	0.000000
ConvLSTM	V	B	0.117884	0.119743	0.001859	0.000000
ConvLSTM	M	A	0.067974	0.068004	0.000030	0.000000
ConvLSTM	M	B	0.074327	0.074646	0.000320	0.000000

sizes 3/3/6, with 21 matched and 45 unmatched pairs. The largest class receives matched endpoint mass $15/21 = 5/7$. Every unmatched pair crosses classes, so its normalized endpoint mass in any one class is at most $1/2$. Thus even unrestricted unmatched references cannot balance the complete family's equal-matched estimand. Adding its two missing histories would not solve that design problem.

For the selected full-support witness we maximize t subject to balance, $\sum_{mr} x_{mr} = 1$ and $x_{mr} \geq t$ on every eligible edge, allowing the matched masses to change. Exact rational primal/dual certificates give minimum joint edge masses $1/12$ for A's ten edges and $1/88$ for B's 66. B uses all twelve matched pairs and all seventeen unmatched pairs with eligible support; the other sixteen unmatched pairs have no compatible matched coefficient. The smallest matched mass in this chosen B witness is $1/44$. This differs from the separate objective maximizing minimum matched-pair mass, whose optimum is $1/20$; those objectives and witnesses are not interchangeable. The selected witness is obtained from the symbolic input graph, and multiple optimal witnesses may exist.

Complete subfamily census. All 54 two-per-class subfamilies are retained in the input table. Eight lack at least one exact reference and are not assigned a complete-comparison ratio. Of the 46 covered, six have balanced default references; two further subfamilies can balance by reweighting references alone. Among the remaining 38 covered candidates, six admit balance only after giving their matched pair from B's four-history class zero mass. These six remain in the complete 46-subfamily default-exact results, but cannot give full-matched-coverage balanced comparisons. Every feasible witness and infeasibility certificate is included in the companion files. Subfamily identifiers in the outcome tables are assigned by sorted condition-name sets; symbolic-program identifiers use a different order and are joined by names, not by number.

The naturally balanced subfamilies are listed below; the table records every eligible balanced construction, without selecting an output direction.

R15 ID	R17 CSV ID	Conditions
S001	S01	E012, E021, E102, E120, E201, E210
S006	S09	E012, E021, E102, E120, L021, L201
S008	S25	E012, E021, E102, E201, L012, L021
S017	S06	E012, E021, E120, E210, L102, L201
S019	S22	E012, E021, E201, E210, L012, L102
S024	S30	E012, E021, L012, L021, L102, L201

All pair/endpoint weights, coefficient contributions, clip-level readouts, complete 46-subfamily aggregates and exact certificates are archived in `supplement/r15`. Subfamilies are correlated reuse of one finite study slice. Neither their count nor a weight optimum is new sample evidence; no confidence interval, equivalence, accuracy gain or selection utility is claimed. Numerical verification details, including floating-point signs near zero, are in the companion reproducibility statement.

Table 48: Complete finite distribution of exact-distance Q ratios across all 46 covered subfamilies, separated by default endpoint balance. Each subfamily ratio uses equal-clip means. Min/median/max summarize the subfamilies; above/below counts compare those ratios to one, not the 32 clip signs. The overlapping subfamilies are not independent replications.

Artifact	Arm	Horizon	Subfamilies	Min	Median	Max	Above/below
SimVP	V	first	Balanced (6)	0.9477	1.0082	1.0276	4/2
SimVP	V	first	Unbalanced (40)	0.8033	0.9266	1.0596	5/35
SimVP	V	mean10	Balanced (6)	0.9264	1.0047	1.0606	3/3
SimVP	V	mean10	Unbalanced (40)	0.8545	0.9285	1.0750	11/29
SimVP	M	first	Balanced (6)	0.8830	1.0137	1.0460	4/2
SimVP	M	first	Unbalanced (40)	0.7981	0.8839	1.0650	3/37
SimVP	M	mean10	Balanced (6)	0.9200	1.0142	1.0392	4/2
SimVP	M	mean10	Unbalanced (40)	0.8673	0.9429	1.0612	3/37
ConvLSTM	V	first	Balanced (6)	0.9807	0.9979	1.0084	3/3
ConvLSTM	V	first	Unbalanced (40)	0.9446	0.9960	1.0640	17/23
ConvLSTM	V	mean10	Balanced (6)	0.9852	0.9927	0.9981	0/6
ConvLSTM	V	mean10	Unbalanced (40)	0.9495	1.0031	1.0560	22/18
ConvLSTM	M	first	Balanced (6)	0.9822	1.0023	1.0141	3/3
ConvLSTM	M	first	Unbalanced (40)	0.9088	0.9968	1.1010	20/20
ConvLSTM	M	mean10	Balanced (6)	0.9836	0.9886	1.0060	1/5
ConvLSTM	M	mean10	Unbalanced (40)	0.8999	0.9974	1.0850	18/22

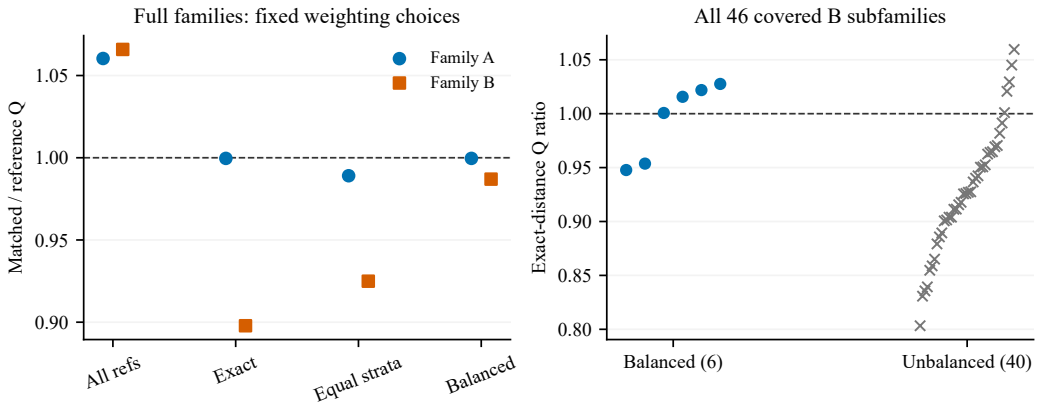

Figure 14: Saved-output sensitivity for SimVP/V/first. Left: four explicitly specified weight schemes on each full family; equal-stratum weights change both means, while the balanced witness changes both in B and coincides with exact weights in A. Right: every covered B subfamily under its default exact references, classified solely by input endpoint balance. Within each balance class, points are ordered by ratio. Points are overlapping subfamilies, not independent samples; dashed lines mark ratio one.

S23 Exact Weight-Sensitivity Ranges and Initial Graph Audit

Finite objectives and certificates. For each of the 576 artifact/arm/family/horizon/readout cells, we take the exact rational average of its 32 saved decimal pair values, then optimize over each of the three prespecified domains. There are 1,728 envelope rows. Each has five final certificates: minimum reference mean, minimum/maximum matched-minus-reference difference, and minimum/maximum ratio, totaling 8,640. Every reference-mean minimum is strictly positive. All horizons are retained; first and step01 duplicate the same forecast readout and are not independent observations.

Writing $x = y + \tau\mathbf{1}$ gives $Ay = b - A\tau\mathbf{1}$, $y \geq 0$. A proposed basis is accepted only after exact verification of all original equalities, nonnegativity, dual reduced costs and primal–dual objective equality. Floating-point arithmetic proposes a basis; it does not determine the saved bound. To minimize a ratio, an exact linear program minimizes $n^\top x - \lambda d^\top x$ and updates $\lambda = n^\top x / d^\top x$. The final certificate has optimum exactly zero. Maximum ratios use the negative objective. Strict denominator positivity then proves global fractional optimality. The certificates concern archived decimal numbers, not the precision of unrounded model arrays or a population parameter.

Closure and positive-support interpretation. Let w be the saved positive feasible witness. For any $x \in \mathcal{P}_0$, $x_\epsilon = (1-\epsilon)x + \epsilon w$ remains balanced and has every edge positive for $0 < \epsilon < 1$. Thus the strictly positive domain is dense in $\mathcal{P}_0$. Since all denominators are bounded away from zero here, continuity makes the closure extrema equal to the strictly positive domain's infimum and supremum. This does not assert that a boundary optimum is attained with positive support. The $t_\star/2$ and $t_\star$ domains are closed, positive-support domains with attained extrema. Across domains, difference and ratio bounds satisfy the expected nesting. A direction category uses a prespecified display band 1 ± 10^{-10}; exact fractions and exact signs relative to one are also retained. The band is not an equivalence threshold.

Table 49: Complete first-frame raw Q ratio ranges on the follow-up outputs. Columns specify the minimum joint edge mass. These are finite weight-sensitivity ranges, not confidence intervals.

Artifact	Arm	Family	0	$t_\star/2$	$t_\star$
SimVP	V	A	0.9653–1.0365	0.9822–1.0177	0.9996
SimVP	M	A	1.0087–1.0284	1.0136–1.0234	1.0185
ConvLSTM	V	A	0.9909–1.0105	0.9957–1.0055	1.0006
ConvLSTM	M	A	0.9962–1.0259	1.0034–1.0183	1.0108
SimVP	V	B	0.9057–1.0614	0.9416–1.0229	0.9843–0.9870
SimVP	M	B	0.8414–1.1142	0.8915–1.0394	0.9637–0.9726
ConvLSTM	V	B	0.9707–1.0164	0.9815–1.0067	0.9935–0.9964
ConvLSTM	M	B	0.9711–1.0254	0.9823–1.0123	0.9960–1.0007

Initial-study graphs. The initial audit reconstructs all 64 clip/arm graphs under three rules, yielding 192 rule-specific graphs. Original weights are uniform over all actual matched pairs and all actual unmatched pairs. Exact rules keep every actual matched pair in the records, assign equal mass to covered matched pairs, and distribute each one's mass uniformly over its available references. Endpoint marginals assign half of a pair's mass to each history. The native history N is recorded with zero pair mass.

In unmasked clip 340, scalar matching adds three edges to the ten-edge coefficient support (13 total); the maximum absolute endpoint-marginal discrepancy is $1/30$. Map matching retains the ten-edge graph and zero discrepancy. In masked clip 324, the actual classes give seven matched pairs, of which four have exact references; three uncovered pairs retain zero weight. Both exact rules have 32 eligible joint edges, compared with 12 coefficient-compatible edges under the actual grouping, and maximum endpoint discrepancy $1/4$. The original all-pair rule in this clip has discrepancy $5/28$. Every other original-rule graph is balanced (32/32 unmasked, 31/32 masked). No degenerate clip is removed. Spatial equality is cross-checked from archived map digests, saved array-equality flags and edge records; no new pixel arrays are read.

The companion directory `supplement/r16` provides all envelope rows, exact pair means, matrices, witnesses and certificates, plus 1,344 history-marginal rows, 2,880 pair-weight rows and 3,651

reconstructed reference edges for the initial audit. These analyses use existing tables only: zero new predictions or training. The Moving MNIST protocol grids total 3,072 forecasts, and the original prospective positive-direction criterion remains unconfirmed. The separate KTH extension adds 334 forecasts (Section S8).

The later ConvLSTM KTH extension adds another 334 scientific forecasts, making 668 KTH forecasts and 3,740 across the described Moving MNIST and KTH grids. These counts exclude earlier development studies and engineering calls.

Table 50: Moving MNIST SimVP, unmasked, first-future raw Q. Complete matched/reference ratio ranges over shared balanced weights. Strictly positive denominators are verified. The range is over fixed outputs, not sampling uncertainty.

Family	Nonnegative closure	Half-floor	Max-min face
A	[0.9653,1.0365]	[0.9822,1.0177]	0.9996
B	[0.9057,1.0614]	[0.9416,1.0229]	[0.9843,0.9870]

S24 Input-Graph Criteria for Endpoint Balance

These statements apply classical finite-dimensional convex geometry (Boyd & Vandenberghe, 2004; Rockafellar, 1970) to the comparison graph. They are not new separation theorems or model-capability results. Let V be a finite history set, M a nonempty finite matched-pair set, R a finite reference-pair set, and $E \subseteq M \times R$ the allowed joint support. Each pair has two distinct endpoints; $a_s \in \{0,1\}^V$ is its incidence vector, with sum two. Write $R_m = \{r : (m,r) \in E\}$. Joint weights $x_{mr} \geq 0$ sum to one. Balance means $\sum_{mr} x_{mr}(a_m - a_r) = 0$; dividing both sides by two gives normalized endpoint marginals.

Proposition 1: fixed matched masses. Fix $u_m \geq 0$ with $\sum_m u_m = 1$. If $u_m > 0$ and R_m is empty, no feasible joint weights exist. Otherwise, put $b = \sum_m u_m a_m$ and

$$C(u) = \sum_{m:u_m>0} u_m \operatorname{conv}\{a_r : r \in R_m\}.$$

Balanced weights with row sums u exist if and only if $b \in C(u)$, equivalently if and only if, for every $f \in \mathbb{R}^V$,

$$f^T b \leq \sum_{m:u_m>0} u_m \max_{r \in R_m} f^T a_r. \tag{14}$$

Proof. For a positive row, x_{mr}/u_m is a conditional probability distribution over its references. Thus its conditional incidence lies in the row's convex hull, and summing over rows gives $b \in C(u)$. Conversely, any such convex-hull representation gives conditional distributions and hence joint weights after multiplication by u_m; zero rows receive zero weights. This proves the first equivalence. The finite Minkowski sum $C(u)$ is nonempty, compact and convex. Its support function is the sum of the row support functions because each row can maximize independently. Membership therefore implies Eq. (14). If $b \notin C(u)$, let z be its nearest point in $C(u)$ and set $f = b - z \neq 0$. For each $y \in C(u)$, expansion of $\|b - z - t(y - z)\|^2$ for $0 \leq t \leq 1$, together with minimality at zero, gives $f^T(y - z) \leq 0$. But $f^T b = f^T z + \|f\|^2$, strictly exceeding the support function. This violates Eq. (14) and proves the converse. Empty zero-weight rows are omitted; no maximum over an empty positive-weight row is defined. □

Proposition 2: flexible weights and positive support. For nonempty E, put $d_e = a_m - a_r$. Balanced probability weights exist exactly when $0 \in \operatorname{conv}\{d_e : e \in E\}$. Weights positive on *every eligible edge* exist exactly when zero lies in the relative interior of this hull. If E is empty, normalized weights do not exist. Edge positivity alone says nothing about a matched row with no eligible edge.

Proof. Normalized balance is precisely a convex combination of the d_e equal to zero. For the positive refinement, a positive combination places zero in the hull. If zero were on its relative boundary, a supporting functional in the affine hull would be nonnegative on all d_e, zero at zero, and strictly positive at some d_e, contradicting the positive combination. Since the affine hull contains zero, it is a

linear space. In the zero-dimensional case the hull is $\{0\}$ and any strictly positive probability vector works. Conversely, if zero is in the relative interior, a small relative ball around it lies in the hull. For each nonzero d_e, choose $\epsilon > 0$ small enough that $-\epsilon d_e$ is in the hull. Combining a representation of this point with mass $\epsilon/(1+\epsilon)$ on d_e represents zero with positive mass on that edge. For a zero vector use unit mass on its edge. Averaging the finitely many representations gives positive mass on every edge, including duplicates. □

When zero lies outside the hull, strict separation yields a potential f with $\min_e f^T d_e > 0$, a finite infeasibility certificate. On a nonempty balanced feasible set, the distinct optima

$$t_\star = \max_x \min_{e \in E} x_e, \qquad \eta_\star = \max_x \min_{m \in M} \sum_{r \in R_m} x_{mr}$$

are attained by compactness. The first is positive exactly when every eligible edge can remain positive; the second is positive exactly when every matched pair can remain positive. The second keeps *all* matched rows, including uncovered rows, whose mass is necessarily zero. Neither criterion guarantees a unique response comparison.

Proposition 3: the unrestricted cross-class case. Suppose V is partitioned into nonempty classes C_i, all matched pairs are within classes, R is *exactly* the set of all cross-class pairs, and $E = M \times R$. No within-class reference is allowed. For fixed u, put $p_v = \frac{1}{2}\sum_m u_m a_m(v)$. Balanced reference weights exist if and only if $\max_i p(C_i) \le \frac{1}{2}$.

Proof. Any cross-class pair contributes at most half its normalized endpoint mass to a class, proving necessity. For sufficiency, place consecutive half-open intervals of lengths $p(C_i)$ on the unit circle, omitting zero-length intervals. Let T be uniform on the circle and take the classes containing T and $T + 1/2$ modulo one. Since every interval has length at most one half, these classes always differ under the half-open convention. Both class marginals are $p(C_i)$, and the coupling is symmetric because the half-rotation is a measure-preserving involution. Conditional on two positive-mass classes, select their histories independently with probabilities $p_v/p(C_i)$. This gives a symmetric ordered cross-class pair distribution with history marginal p. Merging orientations gives an unordered reference distribution q whose normalized endpoint marginal is p. The joint weights $x_{mr} = u_m q_r$ are legal under complete cross-class eligibility, have row sums u, and balance the endpoints. Zero masses require no division by zero. Equality at one half is included. With only one class, no cross-class reference exists and the half-mass condition fails; the construction is not entered. □

For uniform weights on all within-class pairs, $p(C_i) = \binom{n_i}{2} / \sum_j \binom{n_j}{2}$, with positive denominator. Thus the criterion becomes $2\max_i \binom{n_i}{2} \le \sum_j \binom{n_j}{2}$. If eligibility is restricted but references remain cross-class, this is only a necessary condition. For example, take three two-history classes $\{a,b\}, \{c,d\}, \{e,f\}$, uniform matched pairs ab, cd, ef, and allow only reference ac for each row. All class masses are $1/3$, yet the potential equal to one on b, d, e, f and zero on a, c has matched incidence value $4/3$ and reference support value zero, violating Eq. (14).

Finite application checks. The companion retains all 54 subfamilies, the full A/B cases and three previously specified B extensions (B11_C010020, B11_C020010, and B12_full), plus all 192 initial rule/clip/arm graphs. Actual pixel grouping and all uncovered matched pairs remain in the graph. For initial graphs, fixed recorded mass uses the archived covered-pair estimator; fixed uniform mass assigns $1/|M|$ to *every* actual matched pair, a distinct comparison when coverage is partial. Symbolic subfamilies retain all matched pairs uniformly and have no defined default comparison when a row is uncovered. Float simplex proposals are checked with exact rational primal/dual equalities or strict separating certificates. These are finite input-graph checks, not additional predictions or independent test samples.

S25 KTH Execution and Conditional Support

The source videos come from the official KTH action archives. We extract fixed AVI members with HTTP byte ranges, decode in order with OpenCV 4.12, take RGB channel zero, and apply bilinear resizing to float32 before division by 255, as specified by the released loader's channel and resize convention. The AVI-to-image provenance differs from the pre-extracted benchmark PNGs; we

do not claim exact reproduction of the published benchmark score. The checkpoint is the released `kth20_simvp_gsta_one_ep100.pth`. Its single DDP prefix is stripped from keys, with strict key and tensor-value equality checks. The model executes only its first ten-frame block, in FP32 eval mode, batch one, deterministic kernels, and TF32 disabled. Targets are staged separately during decoding and are not read by the model worker.

The six unscorable source clips are person19 jogging; person21 jogging, running and walking; person23 jogging and running (all d1). The scorable counts for subjects 17/19/21/23 are 6/5/3/4. All 24 native predictions and all named construction mappings are retained. Actual-history deduplication saves calls without dropping logical comparisons. On every supported graph, A permits fixed-uniform balance while B does not; their maximum feasible minimum edge masses are respectively 1/12 and 1/88. The analysis uses half these floors, with independently variable per-clip weights. Exact rational primal/dual certificates certify the endpoints for the already computed binary64 pair readouts. Rounding affects presentation only; interval signs are decided before conversion to decimal.

An input-only diagnosis of these failures finds two distinct frames among source frames 0–2 for person19/person21 jogging: the six named A histories reduce to three unique histories and the ten B histories to four. Each family nevertheless has one actual directed-edge class, leaving no unmatched pair. In the other four failed clips, source frames 0–2 are identical and every constructed history within each family is identical. Thus the failure is exact pixel repetition/class merging, not merely a claim of near-periodicity. All six occur in locomotion categories; no causal explanation of that pattern is inferred. This post-run description does not change the input-frozen admission set.

The companion `supplement/r18` contains the frozen protocol, amendments, support ledger, graph certificates, pair and clip readouts, exact envelope certificates, raw-output hash export receipt and independent verification reports. Individual Q, L, D readouts and native errors retain both first-frame and mean-10 horizons. No hypothesis test, capability claim or downstream selection gain is inferred from these finite results.

Table 51: KTH SimVP first-frame signed $Q_M - Q_R$ differences by subject, scaled by 10^5. Each subject row averages its scorable clips; these descriptive cluster summaries do not change the equal-clip aggregate. A default and witness coincide.

Subject	Scorable / 6	A witness	B default	B witness
17	6	12.949	-6.778	1.670
19	5	2.599	-3.110	-0.052
21	3	4.619	-3.335	-0.249
23	4	7.501	-7.007	0.453

Second-architecture execution. The ConvLSTM successor uses the official KTH20 epoch100 checkpoint and exact source commit recorded in `supplement/r19/ASSET_MANIFEST.json`. The adapter preserves official recurrent equations, hidden-state reset and patch layout; no recurrent formula is corrected. A strict state-dictionary load verifies keys and tensor values. Ten observed frames are followed by twenty zero placeholders and an all-zero test scheduled-sampling mask. The official autoregressive model is executed for its full 20-frame interface; only the first ten future outputs are retained and scored. Discarded later outputs cannot affect earlier causal predictions. No native future is passed to inference. Therefore these retained-horizon results do not reproduce the published full $10 \to 20$ benchmark.

The same warm local runtime uses deterministic FP32, batch one and TF32 disabled. Exactly 334 scientific inputs and one all-zero engineering input are consumed once, with no model replay or training. The source and input hashes are bound before dispatch. The companion `supplement/r19` records source acceptance, byte-identical input reuse, raw-output hashes, full Q/L/D pair and clip results, exact rational bounds and independent recount. R18's records remain unchanged.

Table 52: KTH ConvLSTM first-frame signed differences by subject, scaled by 10^5; same conditional subset and equal-clip aggregate as SimVP.

Subject	Scorable / 6	A witness	B default	B witness
17	6	1.736	-2.729	0.266
19	5	0.277	-1.684	-0.465
21	3	0.781	-1.720	-0.289
23	4	0.275	-6.182	-2.023

S26 Post-hoc Step and Subject Sensitivity

These analyses use saved R18/R19 outputs after the primary results and full-paper review. No model calls, training, cohort replacements or primary-readout changes occur. All ten A horizons are included for both models, and all four subject omissions for both families and reported horizons are included. Each retained cohort remains equally weighted by clip, with the same per-clip half-floor domain. A deletion check changes composition; it is neither a new test set nor a population confidence interval.

Table 53: Complete post-hoc A step profiles. Signed-difference ranges are scaled by 10^5; exact signs are given separately because tiny positive endpoints may round to zero. No interval is a confidence interval.

Step	SimVP range	Sign	ConvLSTM range	Sign
1	[6.7286, 8.2214]	+	[0.0233, 1.6707]	+
2	[9.4064, 20.9917]	+	[0.0008, 1.8178]	+
3	[-18.2145, -16.0762]	-	[-0.9721, 0.9607]	$\pm$
4	[10.4818, 20.0851]	+	[-1.8492, 0.2134]	$\pm$
5	[8.0170, 9.8936]	+	[-2.2957, -0.1625]	-
6	[8.2962, 10.6960]	+	[-2.6256, -0.4183]	-
7	[9.5097, 11.9647]	+	[-3.2274, -0.9900]	-
8	[9.5020, 12.1037]	+	[-3.7659, -1.5158]	-
9	[9.0358, 11.8338]	+	[-4.0714, -1.8497]	-
10	[10.7528, 13.2844]	+	[-4.1649, -2.0096]	-

The ten-step mean uses one set of weights per clip shared over the ten times. Its optimized bounds are generally not averages of separately optimized step bounds. For any one feasible weight choice, the mean contrast is the mean of the step contrasts; hence uniformly positive lower bounds at every step cannot yield a negative temporal mean under the same exact readout and domain. The observed negative steps, rather than such an impossible sign pattern, resolve that check. Exact certificates refer to archived binary64 costs; no claim of infinite-precision network arithmetic is made.

Table 54: All 32 post-hoc subject-omission outcomes. $+$ and $-$ mean the entire retained-cohort domain has that sign; $\pm$ means it spans both. Columns omit subjects 17, 19, 21 and 23, retaining respectively 12, 13, 15 and 14 scorable clips. Complete rational endpoints are in the companion JSON.

Model	Family	Horizon	-17	-19	-21	-23
SimVP	A	first	+	+	+	+
SimVP	A	mean 10	+	+	+	+
SimVP	B	first	$\pm$	$\pm$	$\pm$	$\pm$
SimVP	B	mean 10	$\pm$	$\pm$	$\pm$	$\pm$
ConvLSTM	A	first	$\pm$	+	$\pm$	+
ConvLSTM	A	mean 10	-	-	-	+
ConvLSTM	B	first	$\pm$	$\pm$	$\pm$	$\pm$
ConvLSTM	B	mean 10	-	$\pm$	$\pm$	$\pm$

The companion `supplement/r19/posthoc` contains all 360 clip/step objective problems and 720 endpoint certificates, the complete 20 aggregate step cells, with no additional predictions. The independent post-hoc audit and subject-omission JSON are in `supplement/r19`; the latter retains

all 32 exact aggregate intervals. These additions qualify the finite result; they do not alter the original primary claim or establish utility.

S27 Execution and Earlier Evidence Provenance

Reproducible artifact execution. The two public artifacts and minimal adapters are unchanged from the prior replication. SimVP uses the official OpenSTL `mmnist_simvp_s_gsta_one_ep2000.pth` release; the ConvLSTM artifact uses the same release family. Both checkpoints, adapter sources and the runtime are bound by the experiment source manifest. Strict loading verifies every state key and tensor. ConvLSTM uses zero future placeholders and a zero sampling mask; future ground truth is absent from inference. Its supplied cell expression is retained as described in Section S2.

Both studies use FP32, batch size one, evaluation and inference modes, fixed seeds, deterministic algorithms and disabled TF32. The reused local environment is Python 3.12.3 and PyTorch 2.7.1+cu128 on an RTX 5090 D. In the initial study, two separate synthetic smokes precede 448 study forecasts per artifact, and independent recount verifies all 896 predictions. The subsequent prospective distance study separately performs two smokes and 1,088 study forecasts per artifact. Its full raw export precedes acquisition of future targets. Independent implementations verify raw scores, input classes, reference weights and all prespecified readouts. Neither study trains, replays forecasts, stops by response or selects checkpoints. Resource receipts and source/runtime identities accompany the experiment archive; machine-readable tables accompany the paper.

S27.1 Earlier public construction and accounting studies

The initial public slice uses indices0–31 to develop reversed and directed-block constructions. A later fixed slice, indices288–319, evaluates both released artifacts with 1,152 forecasts. Saved-output analyses on that slice separate clipped nonzero-target and zero-target contributions, output excursion, spatial aggregation and temporal cancellation. Those analyses are exploratory because their formulas were specified after aggregate responses were inspected. Region masks describe target intensity, not object identity or motion segmentation.

The subsequent short-block follow-up adds 256 forecasts on those same indices, using P_{102} and P_{201}. It yields unmasked pair-average errors of 0.00976 for SimVP and 0.09098 for ConvLSTM, with native ratios 1.25 and 6.82 and mean absolute aggregate-error differences 0.00145 and 0.01621. Its lower error scale relative to longer blocks motivates the present prospectively fixed pair. Since suffix, omitted frames and temporal placement change together across constructions, this earlier comparison does not isolate a suffix-length effect. The new slice320–351 tests the specified short pair on different clips, not a paired before/after improvement.

S27.2 Controlled RGB development evidence

The earlier simulator study crosses scene identity, yaw and illumination, with paired visible/occluded observations and fixed factual future scoring. Three frozen ConvGRU and three temporal-convolution instances are reused. Actual-input checks show that repeated prefixes can collapse nominal permutations into identical histories; matched early/late placement responses then vary across instances. Original confirmation criteria and exceptions remain in the archived report, including a three-instance criterion met by two ConvGRU instances and a separate cancellation conjunction met by none. Known configurations, overlapping training/validation inputs and discovery-runtime limitations restrict this evidence. It is retained as construction provenance and an input-validation lesson, not an independent generalization result or support for the public-artifact numerical effect.

S27.3 Native-history selection screen and preserved archive

The prespecified development screen uses 128 native clips to choose between the two public artifacts using transition magnitude or its temporal centroid. Both four-cell rules choose SimVP in every cell and give exactly zero development improvement over the global choice. The predeclared development stop leaves confirmation targets and predictions (indices160–287) unevaluated; this round does not

open them. The screen concerns selection from native-history summaries, not a rule conditioned on constructed-block responses. It supplies no selection-utility evidence for the present diagnostic.

The complete pre-R13 report, including all legacy tables, RGB definitions, absolute errors, negative outcomes, qualitative examples and original qualifications, is retained byte-for-byte in the file `pre-R13-complete-history.pdf` under `supplement/legacy/`. Its source and experiment archives remain intact. The condensed provenance here changes presentation, not inclusion in the recorded evidence. The current supplement separates the initial exhaustive grid and its post-hoc analyses from the subsequent prospective distance study. The latter's full machine-readable tables are in `supplement/r14/analysis_tables.zip`.